Rapture of the deep : highs and lows of sparsity in a world of depths

Rémi Gribonval - Inria & ENS Lyon - LIP - Ockham team
remi.gribonval@inria.fr

This project was supported in part by the AllegroAssai ANR project ANR-19-CHIA-0009

Primarily based on joint work with

Quoc-Tung Le

Léon Zheng

Elisa Riccietti

Théo Mary

Antoine Gonon

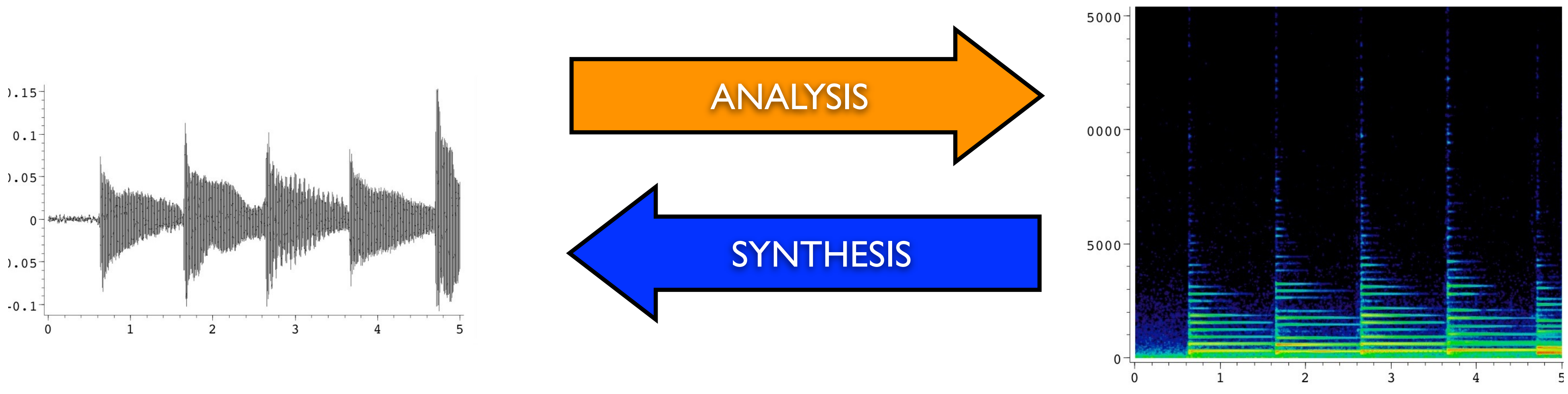
Nicolas Brisebarre

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

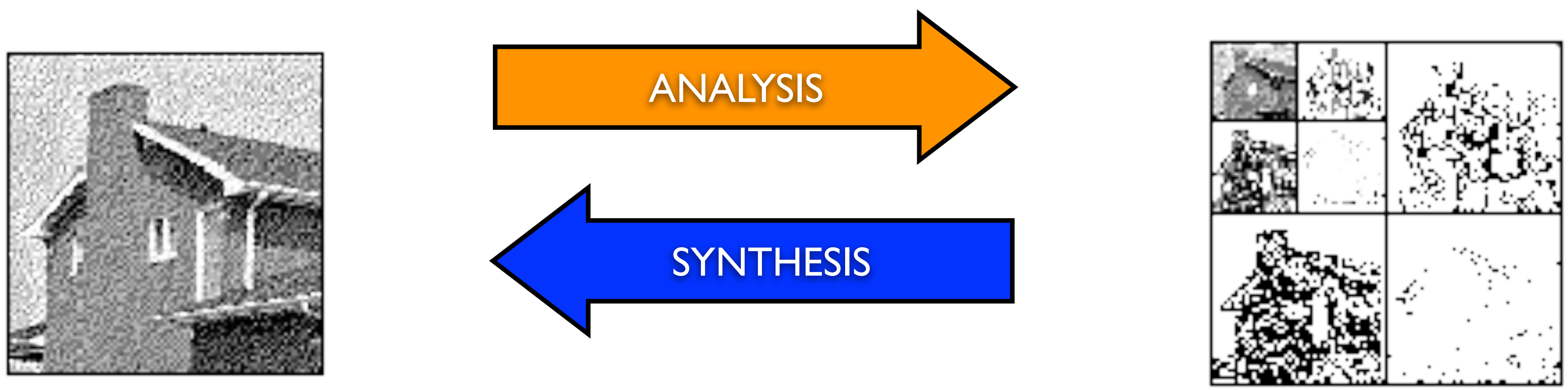
Chaire IA AllegroAssai
ANR-19-CHIA-0009

Sparsity - a gift from nature

- Audio : time-frequency representations

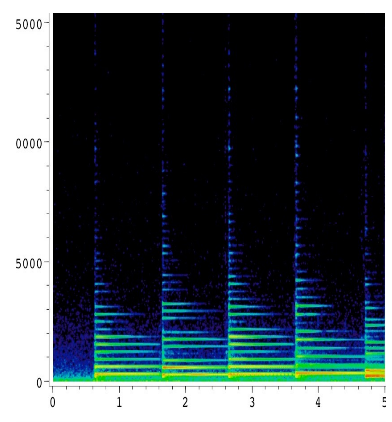


- Images : wavelet transform

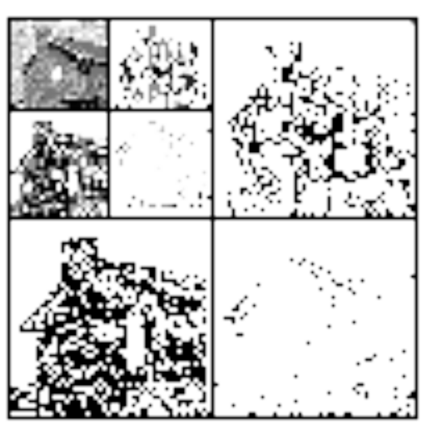


Sparsity & frugality

- **As a natural objective:**
  - bits
  - flops



MP3, AAC



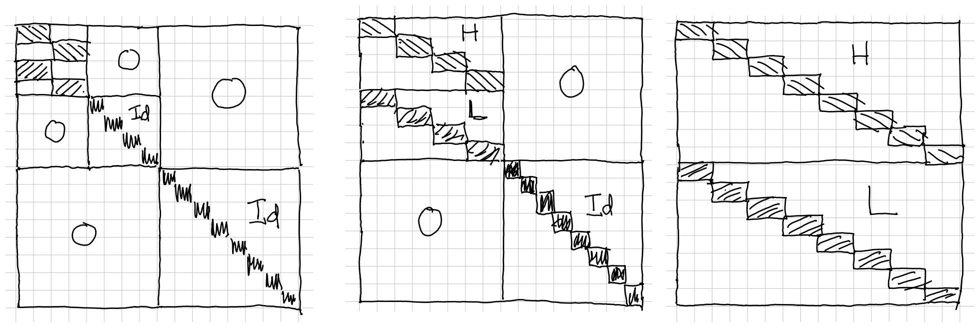
JPEG2000? JPEG !

sparse vector representations

$$\mathbf{z} = \boldsymbol{\Psi}\mathbf{x}$$

Naive matrix multiplication = $O(N^2)$

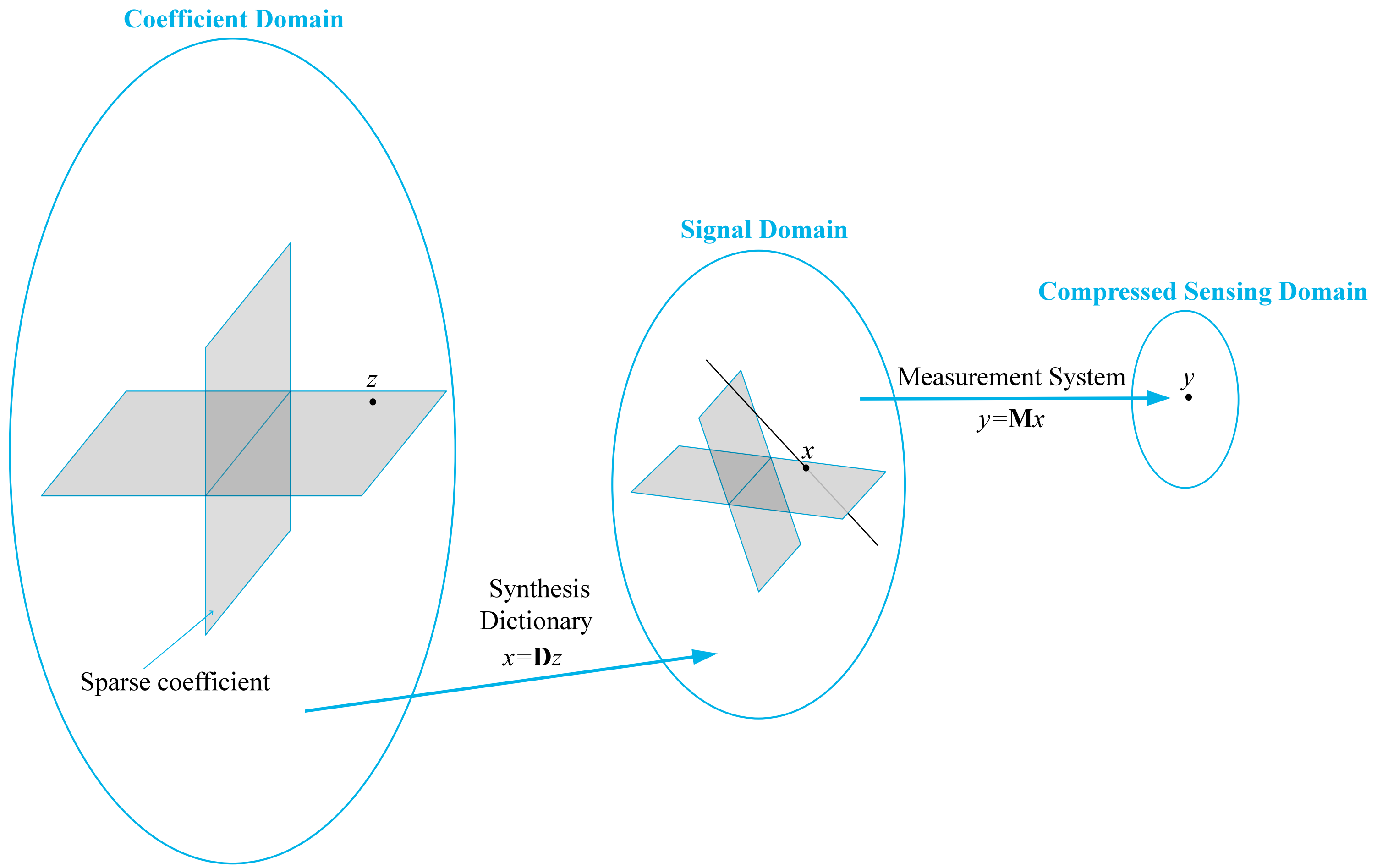
Fast DWT = $O(N)$

$\boldsymbol{\Psi} =$ 

(Deep) sparse matrix product

Sparsity & interpretability

- As a useful ***prior***: identification of latent variables
  - *Prime example =* ***linear inverse problems***



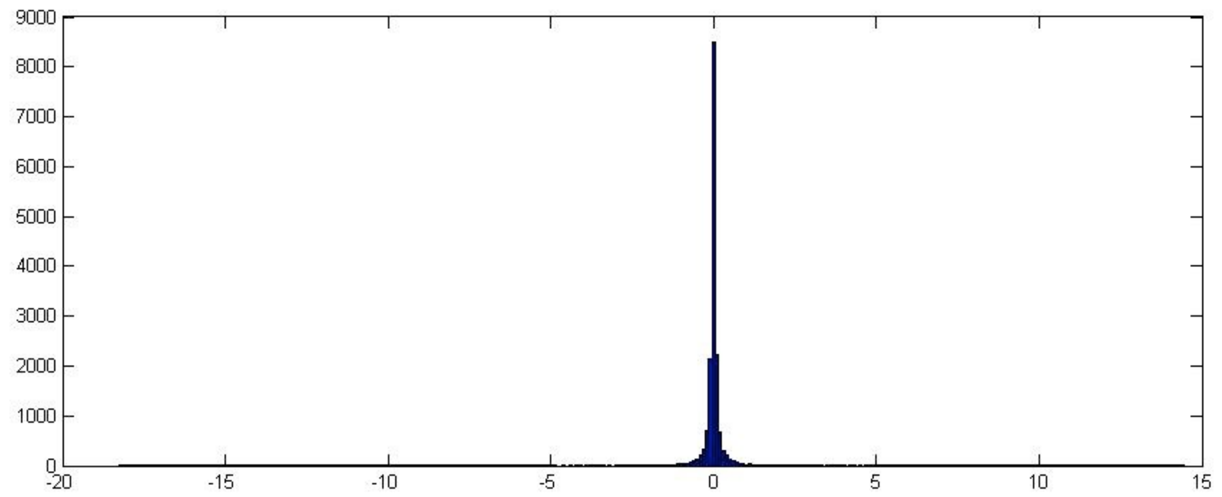
- Provably good algorithms, performance guarantees

Sparse know-how

- L^1 minimization induces sparsity
  - Basis Pursuit / Lasso
- L^2 minimization does not induce sparsity
  - Tikhonov / Ridge regression
- « It’s the support, stupid »
  - Hard = find support (indices of nonzeros)
  - Easy = find coefficients on support (least squares)
- Greed / thresholding is good
  - Matching Pursuits

Deep sparsity ?

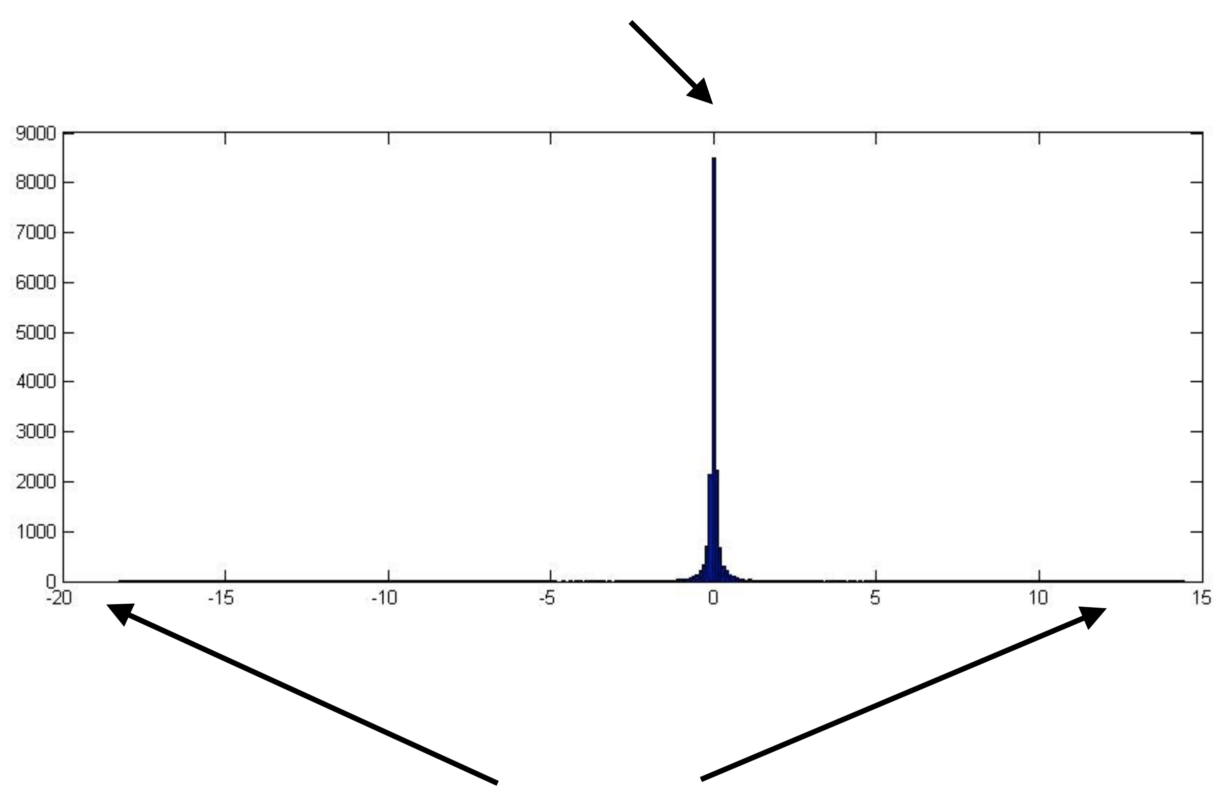
- A gift from nature ?
  - MDCT coefficients of a sound



Deep sparsity ?

- A gift from nature ?
  - MDCT coefficients of a sound

Many small coefficients

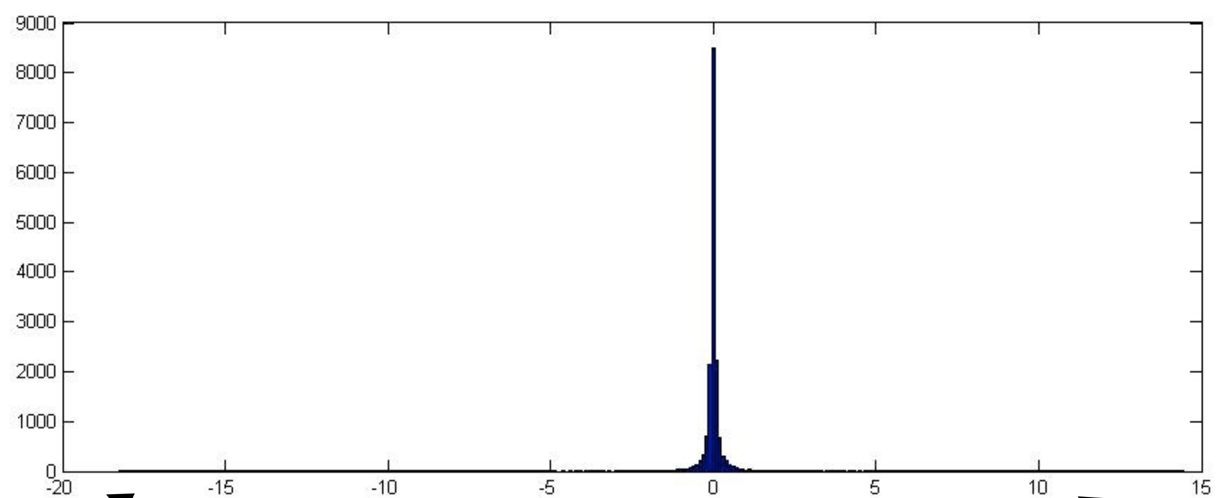


Some (few) coefficients of large magnitude

Deep sparsity ?

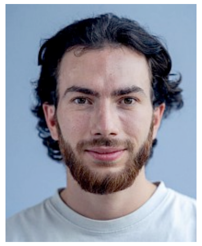
- **A gift from nature ?**
  - MDCT coefficients of a sound

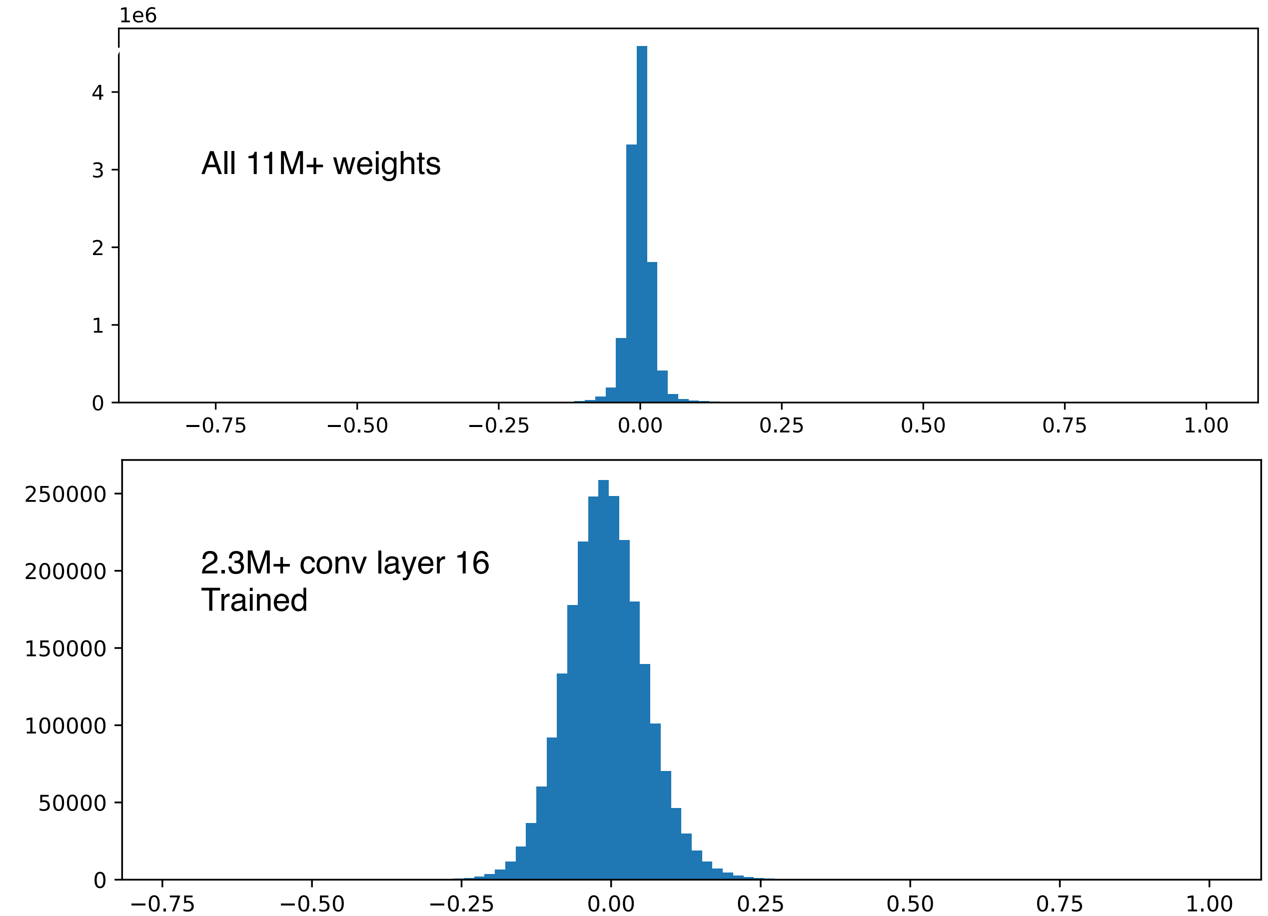
Many small coefficients



Some (few) coefficients of large magnitude

- Weights of a ResNet18 (Trained on ImageNet)
  Thanks to Antoine Gonon





« Deep » sparse know-how ?

- L^1 minimization induces sparsity
  - Basis Pursuit / Lasso
- L^2 minimization does not induce sparsity
  - Tikhonov / Ridge regression
- « It's the support, stupid »
  - Hard = find support (indices of nonzeros)
  - Easy = find coefficients on support (least squares)
- Greed / thresholding is good

Can break down in **deep / multilinear** context

« L1 induces sparsity, but L2 does not » ?

Bilinear sparsity: *blind* deconvolution

- **Goal = recover sparse signals from a convolution**

$$\mathbf{x} = \mathbf{h} \star \mathbf{s}$$

  - « Easy » if one of the signals is known, e.g. with

$$\min \|\mathbf{s}\|_1 \text{ s.t. } \mathbf{x} = \mathbf{h} \star \mathbf{s}$$

  - « Natural » blind approach ?

$$\min \|\mathbf{h}\|_1 + \|\mathbf{s}\|_1 \text{ s.t. } \mathbf{x} = \mathbf{h} \star \mathbf{s}$$

Bilinear sparsity: *blind* deconvolution

- **Goal = recover sparse signals from a convolution**

$$\mathbf{x} = \mathbf{h} \star \mathbf{s}$$

  - « Easy » if one of the signals is known, e.g. with

$$\min \|\mathbf{s}\|_1 \text{ s.t. } \mathbf{x} = \mathbf{h} \star \mathbf{s}$$

  - « Natural » blind approach ?

$$\min \|\mathbf{h}\|_1 + \|\mathbf{s}\|_1 \text{ s.t. } \mathbf{x} = \mathbf{h} \star \mathbf{s}$$

  - Not so simple …

Understanding and evaluating blind deconvolution algorithms

Anat Levin[1,2], Yair Weiss[1,3], Fredo Durand[1], William T. Freeman[1,4]
[1]MIT CSAIL, [2]Weizmann Institute of Science, [3]Hebrew University, [4]Adobe

CVPR 2009

over time [19]. In this paper we analyze the source of the MAP failure. We show that counter-intuitively, the most favorable solution under a sparse prior is usually a blurry image and not a sharp one. Thus, the global optimum of the MAP approach is the no-blur explanation. We discuss so-

A FUNDAMENTAL PITFALL IN BLIND DECONVOLUTION WITH SPARSE AND SHIFT-INVARIANT PRIORS

Alexis Benichoux[1], Emmanuel Vincent[2], Rémi Gribonval[3]

ICASSP 2013

Given a mixture x, we show here that the global minima of (P2) and (P1) are trivial reconstructions, in the sense that the estimated filter is equal to a Dirac pulse. Let δ_0 be the Dirac

$$\mathbf{h}^* = \delta_0, \ \mathbf{s}^* = \mathbf{x}$$

Behind the scene
(in modern words)
scaling ambiguities
~implicit bias
– coming up –

ReLU network training - weight decay

- **Weight decay = quadratic penalty**
  - Shallow ReLU-net implements $f(\mathbf{x}) = \sum_{\text{neurons } i} \mathbf{u}_i \, \mathtt{ReLU}(\mathbf{v}_i^\top \mathbf{x})$
    - (+variant with biases)
  - Training = risk minimization $\min_{\mathbf{u}_i, \mathbf{v}_i} \mathtt{loss}(f) + \sum_i \left(\|\mathbf{u}_i\|_2^2 + \|\mathbf{v}_i\|_2^2 \right)$
  - « Natural » : *non*convex + promotes *flatness* (≠sparsity) ?

ReLU network training - weight decay

- **Weight decay = quadratic penalty**
  - Shallow ReLU-net implements $f(\mathbf{x}) = \sum_{\text{neurons } i} \mathbf{u}_i \, \mathtt{ReLU}(\mathbf{v}_i^\top \mathbf{x})$
    - (+variant with biases)
  - Training = risk minimization $\min_{\mathbf{u}_i, \mathbf{v}_i} \mathtt{loss}(f) + \sum_i \left(\|\mathbf{u}_i\|_2^2 + \|\mathbf{v}_i\|_2^2 \right)$
  - « Natural » : *non*convex + promotes *flatness* (≠sparsity) ?
- **Not so simple …**
  - « Neural balance » at global minimum

$$\|u_i\|_2 = \|v_i\|_2, \ \forall i$$

  - ~Solution to (convex) Group Lasso
    - See e.g.

Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks

Mert Pilanci [1] Tolga Ergen [1]

ICML 2020

Behind the scene
(in modern words)
scaling ambiguities
~implicit bias
– coming up –

Behind the scene

Least Absolute Shrinkage is Equivalent to Quadratic Penalization

Yves Grandvalet ICANN 1998

Blind deconvolution

- Scaling ambiguity

$$(\mathbf{h}/\lambda) \star (\lambda \mathbf{s}), \ \forall \lambda \neq 0$$

- Apparent regularization

$$\|\mathbf{h}\|_1 + \|\mathbf{s}\|_1$$

- Effective regularization

$$\min_{\lambda} \|\mathbf{h}/\lambda\|_1 + \|\lambda \mathbf{s}\|_1 \propto \sqrt{\|\mathbf{h}\|_1 \|\mathbf{s}\|_1}$$

- + a consequence:
  - For every pair such that $\mathbf{x} = \mathbf{h} \star \mathbf{s}$

$$\|\mathbf{x}\|_1 \|\delta_0\|_1 = \|\mathbf{x}\|_1 \leq \|\mathbf{h}\|_1 \|\mathbf{s}\|_1$$

A FUNDAMENTAL PITFALL IN BLIND DECONVOLUTION WITH SPARSE AND SHIFT-INVARIANT PRIORS

Alexis Benichoux[1], Emmanuel Vincent[2], Rémi Gribonval[3]

ICASSP 2013

Behind the scene

Least Absolute Shrinkage is Equivalent to Quadratic Penalization

Yves Grandvalet

ICANN 1998

Blind deconvolution

- Scaling ambiguity

$$(\mathbf{h}/\lambda) \star (\lambda \mathbf{s}),\ \forall \lambda \neq 0$$

- Apparent regularization

$$\|\mathbf{h}\|_1 + \|\mathbf{s}\|_1$$

- Effective regularization

$$\min_{\lambda} \|\mathbf{h}/\lambda\|_1 + \|\lambda \mathbf{s}\|_1 \propto \sqrt{\|\mathbf{h}\|_1 \|\mathbf{s}\|_1}$$

- \+ a consequence:
  - For every pair such that $\mathbf{x} = \mathbf{h} \star \mathbf{s}$

$$\|\mathbf{x}\|_1 \|\delta_0\|_1 = \|\mathbf{x}\|_1 \leq \|\mathbf{h}\|_1 \|\mathbf{s}\|_1$$

A FUNDAMENTAL PITFALL IN BLIND DECONVOLUTION WITH SPARSE AND SHIFT-INVARIANT PRIORS

Alexis Benichoux[1], Emmanuel Vincent[2], Rémi Gribonval[3]

ICASSP 2013

ReLU nets & weight decay

$$(\mathbf{u}/\lambda)\mathtt{ReLU}(\lambda \mathbf{v}^\top \mathbf{x}),\ \forall \lambda > 0, \forall \mathbf{x}$$

$$\sum_i \|\mathbf{u}_i\|_2^2 + \|\mathbf{v}_i\|_2^2$$

$$\sum_i \|\mathbf{u}_i \mathbf{v}_i^\top\|_2$$

- Analog to Group Lasso
- See e.g. survey

IEEE Signal Processing Magazine, 2023

Rahul Parhi and Robert D. Nowak

Deep Learning Meets Sparse Regularization

A signal processing perspective

Greed is good ?

- **Alice**

- **Bob**

- Train & release ReLU net
  - ResNet 18 on ImageNet
- Greedy sparsification
  - Prune-out 50% largest weights on layers 2&3
    - **Top1 ≈32% on ImageNet**
    - **Top5 ≈54%**

!&3

Figures thanks to Antoine Gonon

Greed is good ?

- **Alice**

- **Bob**

- Train & release ReLU net →
  - ResNet 18 on ImageNet

- Greedy sparsification
  - Prune-out 50% largest weights on layers 2&3
    - **Top1 ≈32% on ImageNet**
    - **Top5 ≈54%**

- **4lice**
- with *adversarial* rescaling

- Idem
  - Prune-out 50% largest weights on layers 2&3
    - **Top1 ≈20%**
    - **Top5 ≈40%**

Figures thanks to Antoine Gonon

Greed is good ?

Régularisation implicite des factorisations de faible rang pénalisées

Jérémy E. Cohen
Univ Lyon, INSA-Lyon, UCBL, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100 Villeurbanne, France

- **Alice**

- **Bob**

- Train & release ReLU net →
  - ResNet 18 on ImageNet

- **4lice**
- with *adversarial* rescaling

- Greedy sparsification
  - Prune-out 50% largest weights on layers 2&3
    - **Top1 ≈32% on ImageNet**
    - **Top5 ≈54%**

- Idem
  - Prune-out 50% largest weights on layers 2&3
    - **Top1 ≈20%**
    - **Top5 ≈40%**

Figures thanks to Antoine Gonon

« It’s the support, stupid » … really ?

Optimization with support constraints

Linear inverse problem

$$\min_{x} \|z - \mathbf{M}x\|_2^2$$

under contraint on $\mathtt{supp}(x)$

- Sparsity constraint
  - unknown support
  - NP-hard

Adaptive Greedy Approximations
G. Davis, S. Mallat, and M. Avellaneda
Constr. Approx. (1997) 13: 57–98

- *Fixed* support
  - Used in Greedy / Iterative methods

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 41, NO. 12, DECEMBER 1993 3397
Matching Pursuits With Time-Frequency Dictionaries
Stéphane G. Mallat, *Member, IEEE*, and Zhifeng Zhang

  - Easy = least-squares

Bilinear inverse problem

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{A} - \mathbf{X}\mathbf{Y}\|_F^2$$

under constraints on $\mathtt{supp}(\mathbf{X}), \mathtt{supp}(\mathbf{Y})$

- Sparsity constraint
  - covers sparse PCA
  - NP-hard

[Magdon-Ismael, **NP-hardness and inapproximability of sparse PCA**, *Information Processing Letters*, 2017]

- *Fixed* support
  - E.g. : LU factorization
  - Easy or hard ?

Optimization with support constraints

Linear inverse problem

$$\min_{x} \|z - \mathbf{M}x\|_2^2$$

under contraint on $\mathrm{supp}(x)$

- Sparsity constraint
  - unknown support
  - NP-hard

Adaptive Greedy Approximations

G. Davis, S. Mallat, and M. Avellaneda

Constr. Approx. (1997) 13: 57–98

- *Fixed* support
  - Used in Greedy / Iterative methods

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 41, NO. 12, DECEMBER 1993 3397

Matching Pursuits With Time-Frequency Dictionaries

Stéphane G. Mallat, *Member, IEEE*, and Zhifeng Zhang

  - Easy = least-squares

Bilinear inverse problem

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{A} - \mathbf{X}\mathbf{Y}\|_F^2$$

under constraints on $\mathrm{supp}(\mathbf{X}), \mathrm{supp}(\mathbf{Y})$

- Sparsity constraint
  - covers sparse PCA
  - NP-hard

[Magdon-Ismael, **NP-hardness and inapproximability of sparse PCA**, *Information Processing Letters*, 2017]

- *Fixed* support
  - E.g. : LU factorization
  - Easy or hard ?

Optimization with support constraints

Linear inverse problem

$$\min_{x} \|z - \mathbf{M}x\|_2^2$$

under contraint on $\mathrm{supp}(x)$

- Sparsity constraint
  - unknown support
  - NP-hard

Adaptive Greedy Approximations

G. Davis, S. Mallat, and M. Avellaneda

Constr. Approx. (1997) 13: 57–98

- ***Fixed*** support
  - Used in Greedy / Iterative methods

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 41, NO. 12, DECEMBER 1993 3397

Matching Pursuits With Time-Frequency Dictionaries

Stéphane G. Mallat, *Member, IEEE*, and Zhifeng Zhang

  - Easy = least-squares

Bilinear inverse problem

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{A} - \mathbf{X}\mathbf{Y}\|_F^2$$

under constraints on $\mathrm{supp}(\mathbf{X}), \mathrm{supp}(\mathbf{Y})$

- Sparsity constraint
  - covers sparse PCA
  - NP-hard

[Magdon-Ismael, **NP-hardness and inapproximability of sparse PCA**, *Information Processing Letters*, 2017]

- ***Fixed*** support
  - E.g. : LU factorization
  - Easy or hard ?

Curses of fixed support factorization

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{A} - \mathbf{XY}\|_F^2$$

With « **fixed** » **support** $\mathrm{supp}(\mathbf{X}) \subset I, \mathrm{supp}(\mathbf{Y}) \subset J$

Quoc-Tung Le

Elisa RicciETTI

- **NP-hard to approximate**
- **Minimizer does not always exist**
  - Similar to low-rank tensor decompositions … but also LU factorization
  - Requires regularization (e.g. « weight decay » $\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2$)
  - Phenomenon extends to shallow ReLU nets

Details in:

SPURIOUS VALLEYS, NP-HARDNESS, AND TRACTABILITY OF SPARSE MATRIX FACTORIZATION WITH FIXED SUPPORT

QUOC-TUNG LE*, ELISA RICCIETTI*, AND REMI GRIBONVAL*

SIMAX 2023

Does a sparse ReLU network training problem always admit an optimum?

Quoc-Tung Le Elisa Riccietti Rémi Gribonval

NeurIPS2023

Curses of fixed support factorization

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{A} - \mathbf{XY}\|_F^2$$

With « **fixed** » **support** $\mathrm{supp}(\mathbf{X}) \subset I, \mathrm{supp}(\mathbf{Y}) \subset J$

Quoc-Tung Le

Elisa Ricciietti

- **NP-hard to approximate**
- **Minimizer does not always exist**
  - Similar to low-rank tensor decompositions … but also LU factorization
  - Requires regularization (e.g. « weight decay » $\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2$)
  - Phenomenon extends to shallow ReLU nets
- **Is it really a problem ?**

Details in:

SPURIOUS VALLEYS, NP-HARDNESS, AND TRACTABILITY OF SPARSE MATRIX FACTORIZATION WITH FIXED SUPPORT

QUOC-TUNG LE*, ELISA RICCIETTI*, AND REMI GRIBONVAL*

SIMAX 2023

Does a sparse ReLU network training problem always admit an optimum?

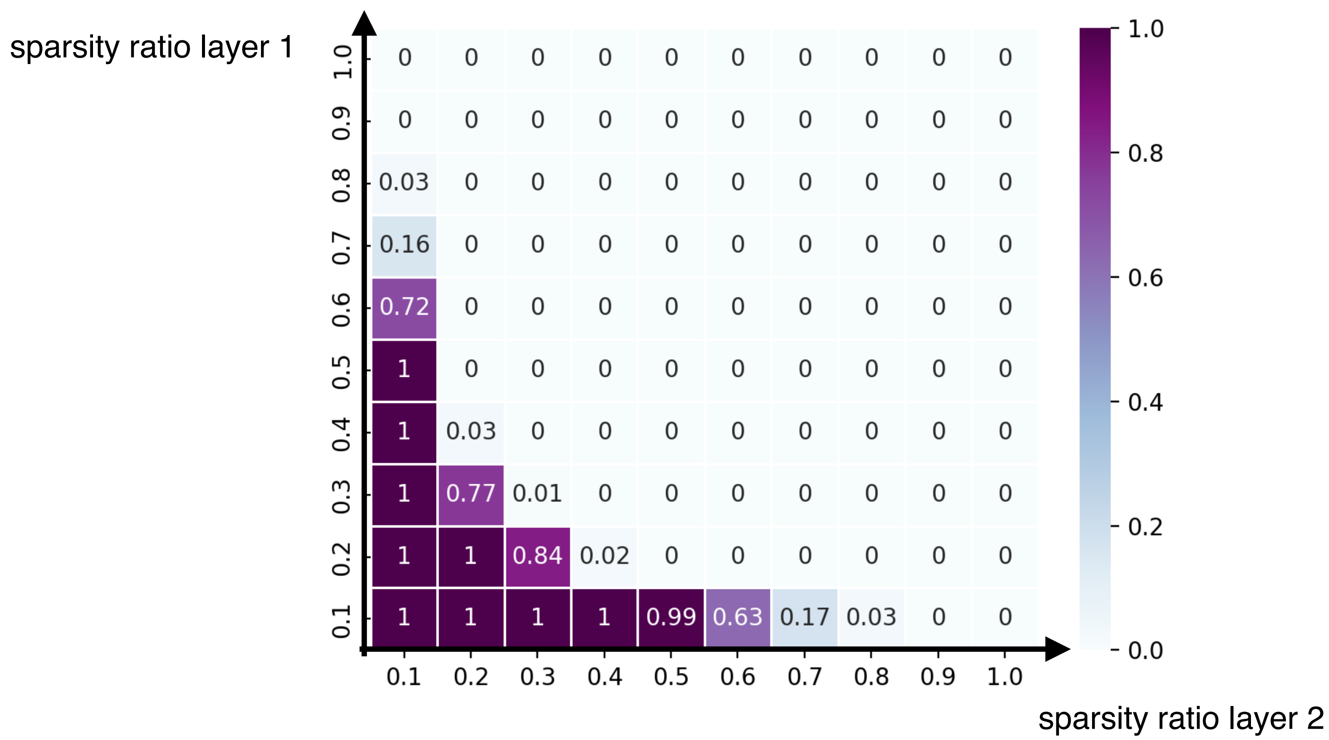
Quoc-Tung Le Elisa Riccietti Rémi Gribonval

NeurIPS2023

Do bad supports occur in practice ?

Experiment

- Shallow ReLU network $f_\theta(x) = \mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1 x)$
- Random support constraints with various sparsities
- Algorithm to detect bad supports
  - Guaranteed: no false positive (some false negatives are possible)
  - Empirical probability of bad support

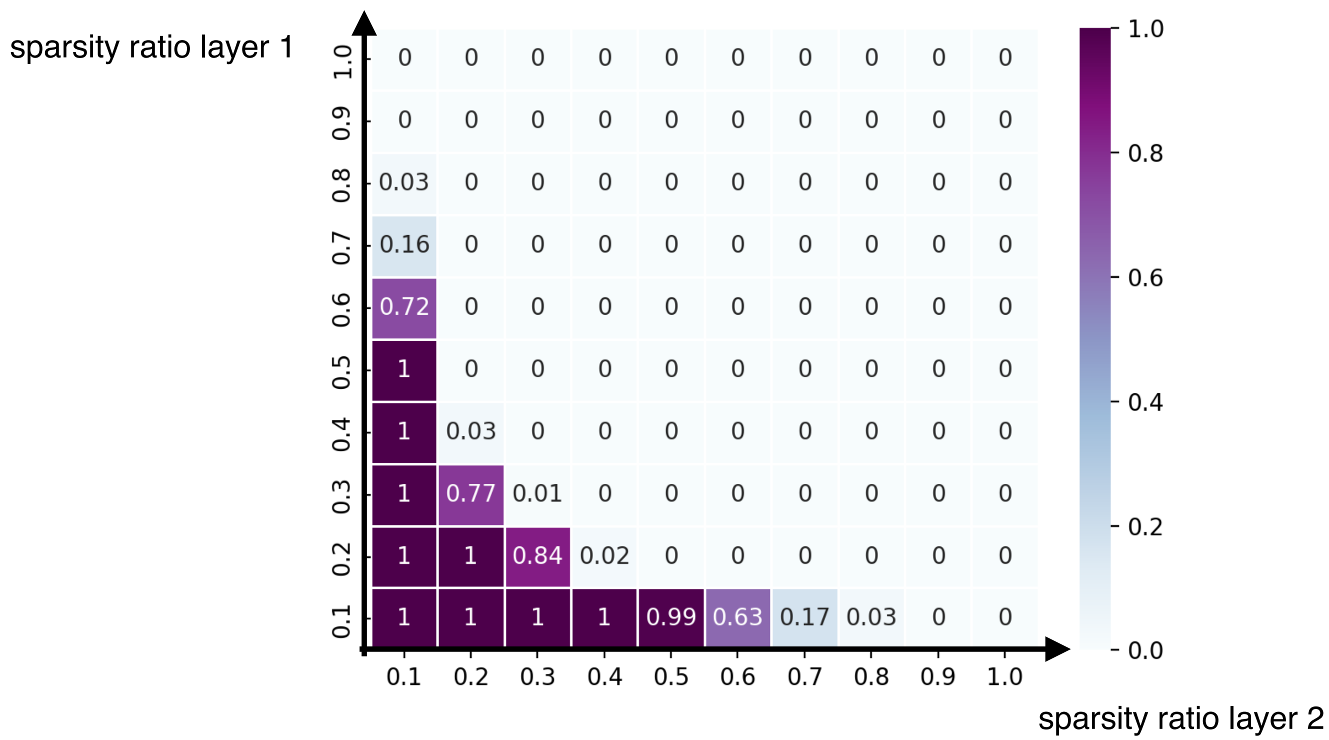


$$\text{sparsity ratio} = \frac{\#\text{nonzero entries}}{\#\text{total nb entries}}$$

Do bad supports occur in practice ?

Experiment

- Shallow ReLU network $f_\theta(x) = \mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1 x)$
- Random support constraints with various sparsities
- Algorithm to detect bad supports
  - Guaranteed: no false positive (some false negatives are possible)
  - Empirical probability of bad support



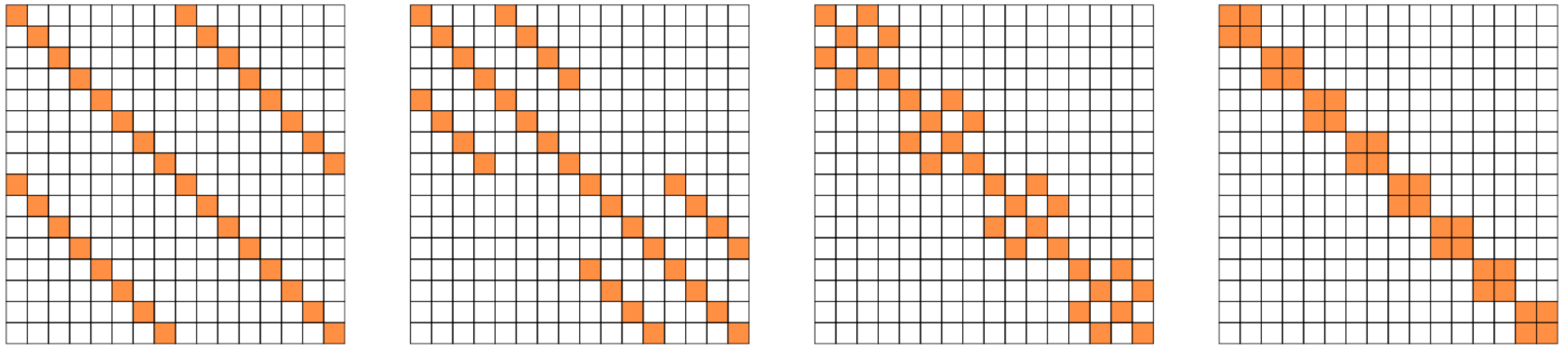
sparsity ratio = $\dfrac{\#\text{nonzero entries}}{\#\text{total nb entries}}$

How to ensure "nice" supports ?

Promises of *butterfly* supports

"*Butterfly*" supports

Ubiquitous for fast *linear* transforms (FFT ...)



(a) $\mathbf{S}_{\mathtt{bf}}^{(1)}$ (b) $\mathbf{S}_{\mathtt{bf}}^{(2)}$ (c) $\mathbf{S}_{\mathtt{bf}}^{(3)}$ (d) $\mathbf{S}_{\mathtt{bf}}^{(4)}$

- solid roots in numerical analysis (multipole methods, hierarchical matrices)
- efficient $O(n \log n)$ implementation instead of $O(n^2)$ + expressivity

Rapidly emerging for (*nonlinear*) deep learning

- see e.g. [T. Dao & al, Monarch: Expressive Structured Matrices for Efficient and Accurate Training, ICML, 2022]

Recent results on *linear* butterflies


Quoc-Tung Le

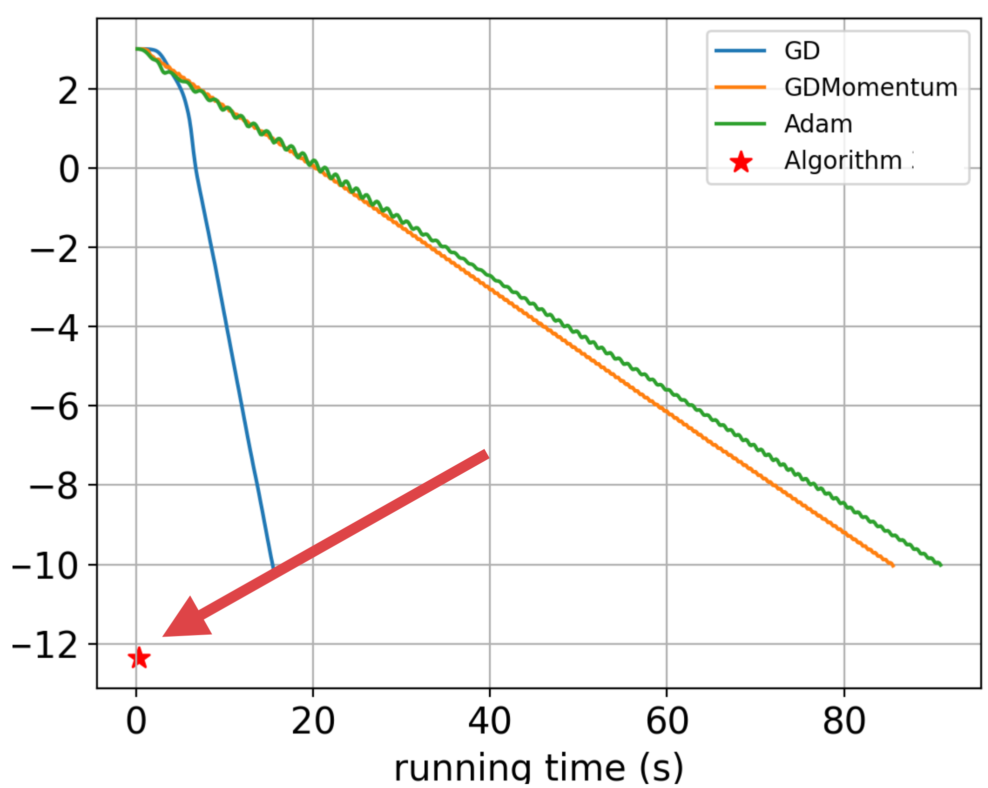

Léon Zheng


Elisa Riccietti

Hierarchical factorization alg.

- **Input:** a (dense) matrix $\mathbf{A}$
- **Output:** butterfly factors
- **Approach:** one factor at a time
  - ***exploiting structure of sparsity pattern***
- **Results**
  - ***Existence*** + ***Identifiability*** of optimum
    - (+ongoing : stability to noise)
  - ***Frugality of hierarchical approach***:
    - efficiency >> gradient descent

log(approximation error) - with two factors



Recent results on *linear* butterflies

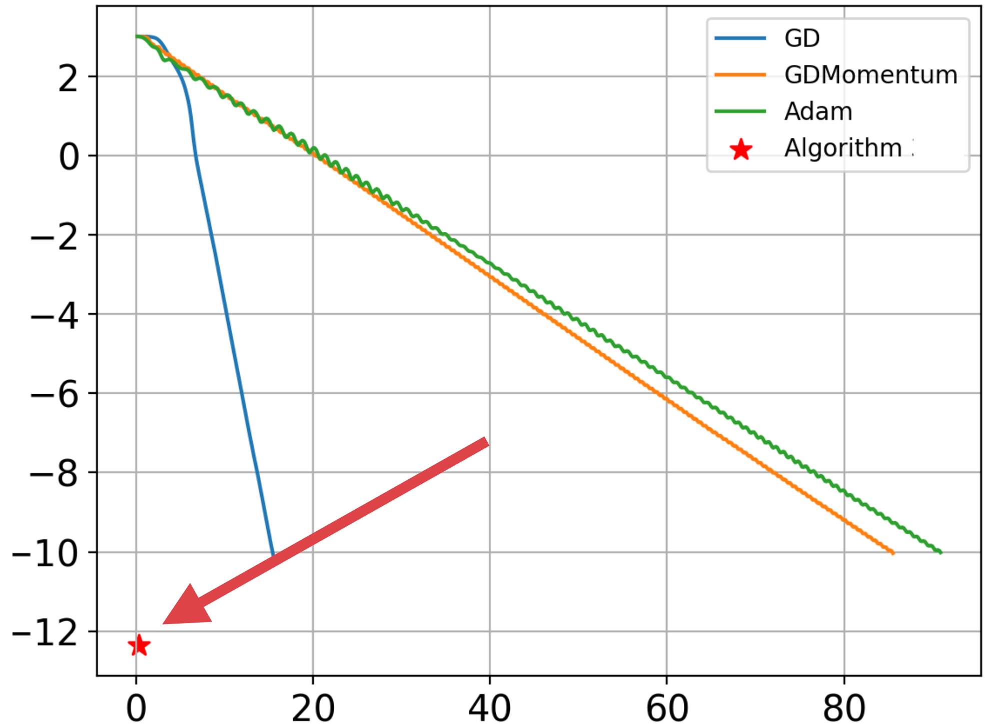
Elisa Riccietti

Théo Mary

Hierarchical factorization alg.

- **Input:** a (dense) matrix $\mathbf{A}$
- **Output:** butterfly factors
- **Approach:** one factor at a time
  - ***exploiting structure of sparsity pattern***
- **Results**
  - ***Existence*** + ***Identifiability*** of optimum
    - (+ongoing : stability to noise)
  - ***Frugality of hierarchical approach***:
    - efficiency >> gradient descent

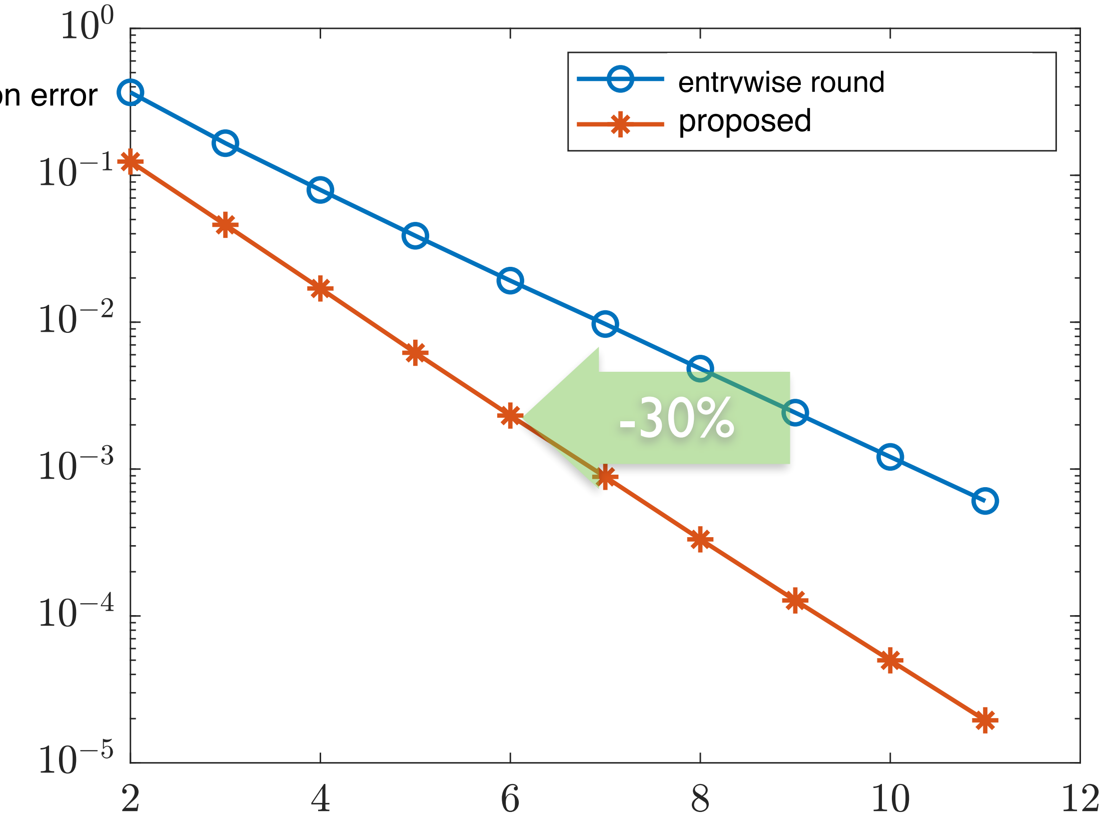
log(approximation error) - with two factors



Efficient quantization alg.

- **Input:** butterfly factors
- **Output:** *quantized* butterfly factors
- **Approach:** one factor at a time
  - ***exploiting rescaling invariances***
- **Results**
  - **Optimal algorithm** for two factors
    - heuristic for multiple factors
  - **-30% bits** compared to usual rounding

relative quantization error



#bits of mantissa per coefficient

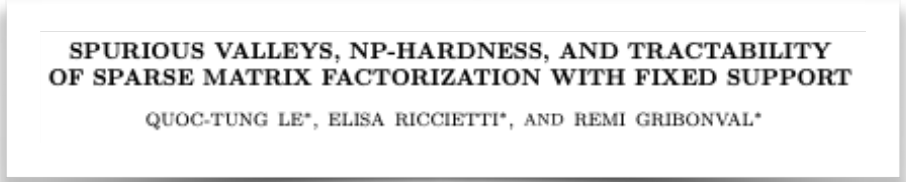
Outlook

Outlook (1)

- **Deep sparsity attractive *objective* for frugal AI ...**
  - Interpretability & privacy
  - Expected flexible resource-performance tradeoffs

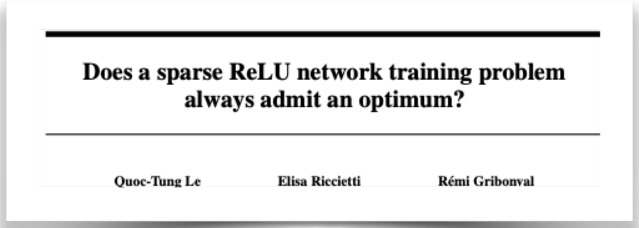
Outlook (1)

- **Deep sparsity attractive *objective* for frugal AI ...**
  - Interpretability & privacy
  - Expected flexible resource-performance tradeoffs
- **... but sparse know-how *breaks down* with depth**
  - No global optimum
  - NP-hardness



SPURIOUS VALLEYS, NP-HARDNESS, AND TRACTABILITY OF SPARSE MATRIX FACTORIZATION WITH FIXED SUPPORT

QUOC-TUNG LE*, ELISA RICCIETTI*, AND REMI GRIBONVAL*



Does a sparse ReLU network training problem always admit an optimum?

Quoc-Tung Le, Elisa Riccietti, Rémi Gribonval

  - +Mismatch between explicit & implicit regularization

Outlook (1)

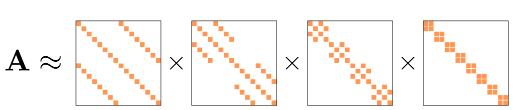
- **Deep sparsity attractive *objective* for frugal AI ...**
  - Interpretability & privacy
  - Expected flexible resource-performance tradeoffs
- **... but sparse know-how *breaks down* with depth**
  - No global optimum
  - NP-hardness

SPURIOUS VALLEYS, NP-HARDNESS, AND TRACTABILITY OF SPARSE MATRIX FACTORIZATION WITH FIXED SUPPORT
QUOC-TUNG LE*, ELISA RICCIETTI*, AND REMI GRIBONVAL*

Does a sparse ReLU network training problem always admit an optimum?
Quoc-Tung Le Elisa Riccietti Rémi Gribonval

  - +Mismatch between explicit & implicit regularization

- **Promises & challenges of butterflies**



  - Stable and tractable learning
  - Improved coarse quantization
  - Ongoing work
    - approximation guarantees
    - improved GPU-friendliness ?
    - performance *vs* efficiency tradeoffs within deep nets ?

Efficient Identification of Butterfly Sparse Matrix Factorizations*
Léon Zheng‡†, Elisa Riccietti†, and Rémi Gribonval†

FAST LEARNING OF FAST TRANSFORMS, WITH GUARANTEES
Quoc-Tung Le*†, Léon Zheng*†◇, Elisa Riccietti†, Rémi Gribonval†

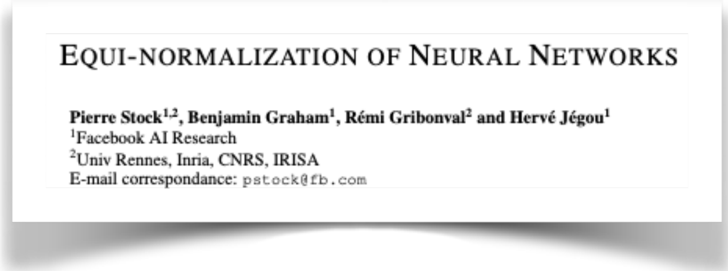
OPTIMAL QUANTIZATION OF RANK-ONE MATRICES IN FLOATING-POINT ARITHMETIC—WITH APPLICATIONS TO BUTTERFLY FACTORIZATIONS*
RÉMI GRIBONVAL†, THEO MARY‡, AND ELISA RICCIETTI†

Outlook (2)

- **Harnessing rescaling-invariance in deep nets ?**
  - Ubiquitous property: ReLU, maxpool, average pool, residual connexions ...

Outlook (2)

- **Harnessing rescaling-invariance in deep nets ?**
  - Ubiquitous property: ReLU, maxpool, average pool, residual connexions ...
- **Empirically: wander in equivalence class**
  - Improved SGD training
  - Reused for quantization by [Nagel & al]


EQUI-NORMALIZATION OF NEURAL NETWORKS

Pierre Stock[1,2], Benjamin Graham[1], Rémi Gribonval[2] and Hervé Jégou[1]
[1]Facebook AI Research
[2]Univ Rennes, Inria, CNRS, IRISA
E-mail correspondance: pstock@fb.com

Outlook (2)

- **Harnessing rescaling-invariance in deep nets ?**
  - Ubiquitous property: ReLU, maxpool, average pool, residual connexions ...
- **Empirically: wander in equivalence class**
  - Improved SGD training
  - Reused for quantization by [Nagel & al]
- **Path-embedding of weights/biases**
  - Rescaling-invariant embedding
    - valid for general DAG architecture
    - combinatorial dimension (#paths)
  - Some consequences
    - **identifiability** analysis
    - statistical **generalization** bounds
    - + **conservation laws** during training
  - Ongoing work
    - provably good/optimal pruning and quantization ?
    - from theoretical tool to computational one ?

EQUI-NORMALIZATION OF NEURAL NETWORKS

Pierre Stock[1,2], Benjamin Graham[1], Rémi Gribonval[2] and Hervé Jégou[1]
[1]Facebook AI Research
[2]Univ Rennes, Inria, CNRS, IRISA
E-mail correspondance: pstock@fb.com

An Embedding of ReLU Networks and an Analysis of Their Identifiability

Pierre Stock & Rémi Gribonval

Constructive Approximation 57, 853–899 (2023)

A PATH-NORM TOOLKIT FOR MODERN NETWORKS: CONSEQUENCES, PROMISES AND CHALLENGES

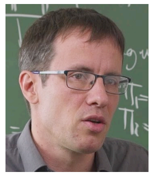
Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti & Rémi Gribonval
Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP, F-69342, LYON Cedex 07, France

Abide by the Law and Follow the Flow: Conservation Laws for Gradient Flows

Sibylle Marcotte
ENS - PSL Univ.
sibylle.marcotte@ens.fr

Rémi Gribonval
Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP,
remi.gribonval@inria.fr

Gabriel Peyré
CNRS, ENS - PSL Univ.
gabriel.peyre@ens.fr



Thank you for your attention !

- Direct contributors

Quoc-Tung Le

Léon Zheng

Elisa Riccietti

Théo Mary

Antoine Gonon

Nicolas Brisebarre

- special thanks to the whole Ockham team @LIP, ENS de Lyon, and particularly to Pascal Carrivain, Simon Delamare, Hakim Hadj-Djilani for their work on `pyfaust & lazylinop` (python+matlab libraries)

Chaire IA AllegroAssai
ANR-19-CHIA-0009