Structure exploiting hierarchical optimization methods

Elisa Riccietti

LIP-ENS Lyon

Collaboration

Guillaume LAUGA

Nelly PUSTELNIK

Paulo GONCALVES

Serge GRATTON

Philippe TOINT

Valentin MERCIER

UNIVERSITÉ DE LYON

ENS DE LYON

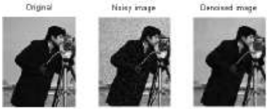
The context

The problem: large scale optimization problems

$$\min_{x \in \mathbb{R}^n} f(x), \quad n \text{ large}$$

- Image restoration

$$\min_x \|Ax - b\|^2 + \lambda \|Dx\|^2$$

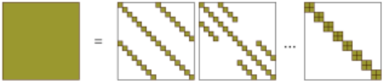


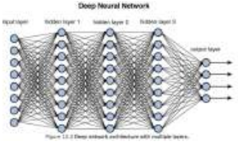
- Neural networks training

$$\min_\theta \sum_{i=1}^{N} \ell(NN(\theta; x_i), y_i)$$

- Matrix factorization

$$\min_{X_1, \dots, X_L} \|A - X_1 \dots X_L\|_F^2$$





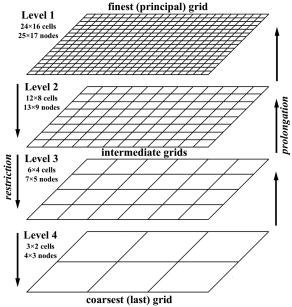
Dimensionality reduction

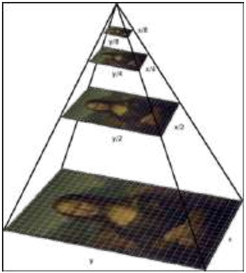
AIM? Reduce the computational cost of the solution process
HOW? Exploit the structure to build approximated subproblems

Dimensionality reduction

AIM? Reduce the computational cost of the solution process

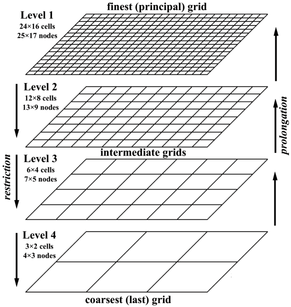
HOW? Exploit the structure to build approximated subproblems

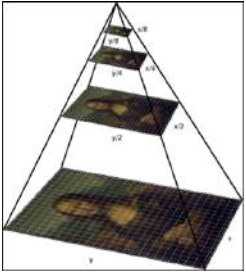




Dimensionality reduction

AIM? Reduce the computational cost of the solution process
HOW? Exploit the structure to build approximated subproblems





Multilevel methods (ML)

Outline

Origins of multilevel methods: multigrid (MG)

A block coordinate descent perspective on MG methods

Application domains of the ML framework

- Physics informed neural networks (distributed)
- Image restoration (hierarchical)

Outline

Origins of multilevel methods: multigrid (MG)

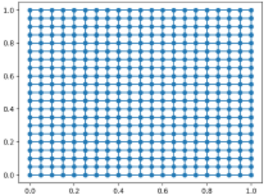
A block coordinate descent perspective on MG methods

Application domains of the ML framework

- Physics informed neural networks (distributed)
- Image restoration (hierarchical)

The numerical solution of PDEs

- Classically PDEs are discretized on a grid
- The resulting linear system $Au = f$ is solved using a fixed point method
- The size of the grids impacts the size of the system and the accuracy of the solution approximation



Fixed point: reduction of the error

Fixed point scheme :

$$u^{(m+1)} = Bu^{(m)} + g$$

After M iterations:

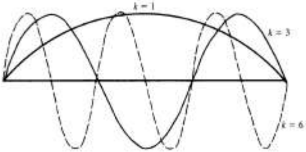
$$e^{(M)} = u^{(M)} - u^* = \sum_{k=1}^{n-1} c_k \lambda_k^M(B) v_k$$

Fourier modes:

$$v_k(j) = \sin\left(\frac{kj\pi}{n}\right), \ k \text{ frequency component}$$

On a n-point grid:

- $1 \leq k < \frac{n}{2}$ low
- $\frac{n}{2} \leq k < n-1$ high

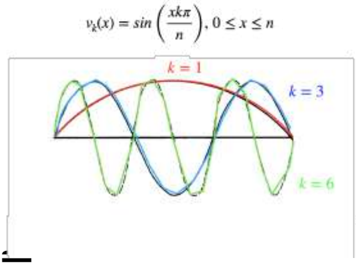


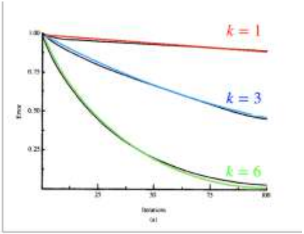
Limitation of iterative schemes: the smoothing property

$$e^{(M)} = u^{(M)} - u^* = \sum_{k=1}^{n-1} c_k \lambda_k^M(B) v_k$$

- $\lambda_1(B) \approx 1$
- $|\lambda_k(B)| < 1/3$ for $n/2 \leq k \leq n-1$

Hard to reduce the low frequency components of the error



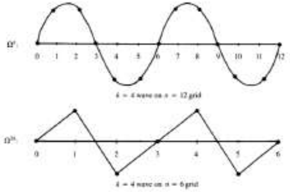


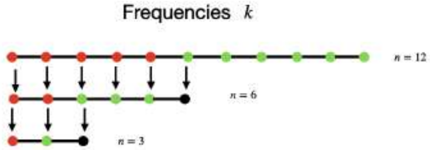
How to make the methods efficient on all frequencies?

Frequency shift!

- **Fine** grid Ω^h with n points: $1 \leq k \leq n-1$
- **Coarse** grid Ω^{2h} with $n/2$ points: $1 \leq k \leq n/2$

Property: $v_k^h(2j) = v_k^{2h}(j)$
Frequency $1 \leq k \leq n/2$ in Ω^h $\rightarrow$ Frequency k in Ω^{2h}





Two-level multigrid methods

Consider a PDE:

$$\mathcal{A}(u) = f.$$

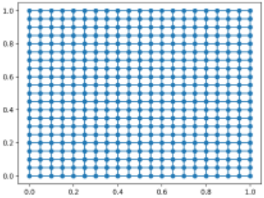
Consider two discretizations:

- Fine grid: $\mathcal{A}_h(u_h) = f_h$
- Coarse grid: $\mathcal{A}_H(u_H) = f_H$

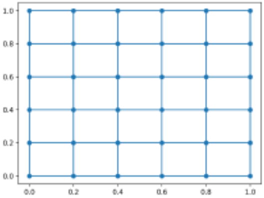
Idea: write the solution u as the sum of a fine and a coarse term:

$$u \sim \underbrace{v_h}_{\in \mathbb{R}^h} + P(\underbrace{e_H}_{\in \mathbb{R}^H}), \; H < h.$$

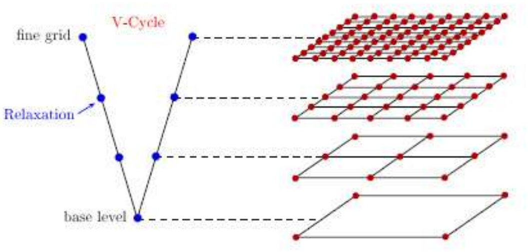
and update the two components in an alternate fashion.



$R \Downarrow$ $\qquad$ $P \Uparrow$



General multigrid methods



W. Briggs, V. Henson, S. McCormick. A Multigrid Tutorial, SIAM, 2000.

Outline

Origins of multilevel methods: multigrid (MG)

A block coordinate descent perspective on MG methods

Application domains of the ML framework
- Physics informed neural networks (distributed)
- Image restoration (hierarchical)

ENS DE LYON

ML methods: abstraction from the PDE context

Problem: $\mathcal{F}$ space of continuous functions parametrized by x

$$\min_{y \in \mathcal{F}} f(y)$$

Approach: we look for y as the sum of two terms

$$y(x) = y_1(x_1) + y_2(x_2).$$

This yields the optimization problem

$$\min_{(x_1, x_2) \in \mathbb{R}^n} f(y_1(x_1) + y_2(x_2)),$$

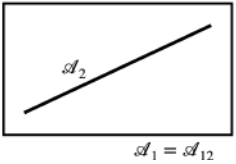
where $n = n_1 + n_2$.

ML methods: approximation spaces

$$\mathcal{A}_{12} = \left\{ y \in \mathcal{F} \mid y(x) = y_1(x_1) + y_2(x_2) \ \text{ for some } \ (x_1, x_2) \in \mathbb{R}^n \right\}$$
$$\mathcal{A}_i = \left\{ y \in \mathcal{F} \mid y(x) = y_i(x_i) \ \text{ for some } \ x_i \in \mathbb{R}^{n_i} \right\} \quad (i = 1, 2).$$

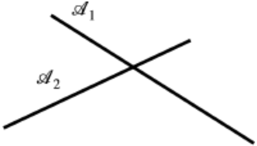
Hierarchical context

$\mathcal{A}_2 \subset \mathcal{A}_1 = \mathcal{A}_{12}$



Distributed context

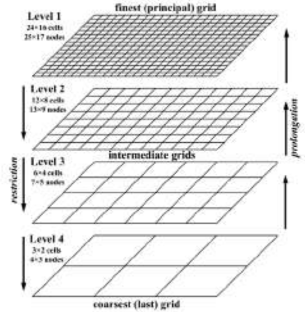
$\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{A}_{12}$



The hierarchical context

Example: classical MG

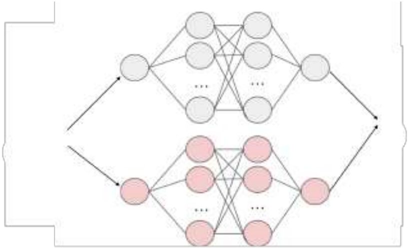
$$f(x) = \frac{1}{2}x^T A x + x^T b$$



The distributed context

Example: neural networks

$$f(x) = \text{loss}$$
$$y_1(x_1) = NN_1(x_1)$$
$$y_2(x_2) = NN_2(x_2)$$



A block coordinate descent (BCD) perspective on ML

The hierarchical context

Alternate:

$$\min_{(x_1, x_2) \in \mathbb{R}^{n_1+n_2}} f(y_1(x_1)+y_2(x_2))$$

and

$$\min_{\substack{x_2 \in \mathbb{R}^{n_2} \\ x_1 \text{ fixed}}} f(y_1(x_1)+y_2(x_2))$$

The distributed context

Alternate:

$$\min_{\substack{x_2 \in \mathbb{R}^{n_1} \\ x_1 \text{ fixed}}} f(y_1(x_1)+y_2(x_2))$$

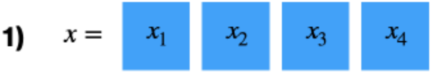
and

$$\min_{\substack{x_1 \in \mathbb{R}^{n_1} \\ x_2 \text{ fixed}}} f(y_1(x_1)+y_2(x_2)).$$

A BCD-ML algorithm: an iteration

How to update x?

1 Partition x in blocks: $(x_1, \ldots, x_n)$



A BCD-ML algorithm: an iteration

How to update x?

2 Select a block i $(x_1, \ldots, x_i, \ldots, x_n)$
- Criterion: $\|\nabla_i f(x)\| \geq \tau \|\nabla f(x)\|$, $\tau \in (0, 1)$

2)

ENS DE LYON

A BCD-ML algorithm: an iteration

How to update x?

3 Update the block:

- p_k iterations of a first-order method (possibly *stochastic*)

$$\min_{x_i} f(x_1, \ldots, x_i, \ldots, x_n) \to x_i^{new}$$

- $x_i \leftarrow x_i^{new}$

3)

$$\min_{x_3} f(x_1, x_2, x_3, x_4)$$

$$x_3 \leftarrow x_3^{new}$$

BCD theory for nonconvex problems

- Powell (1973): cyclic BCD may fail on nonconvex continuously differentiable functions.
- Bertsekas (1999): convergence of cyclic BCD if minimizer along any coordinate direction from any point is unique
- Attouch et all. (2010) + Bolte et all (2014), proximal alternating methods under Kurdyka-Lojasiewicz (KL) property convergence of sequence to stationary points
- Amaral et all. (2022) high (p)-order BCD smooth nonconvex for Lipschitz continuous $\nabla f(x_k)$ + regularized models $\rightarrow O(\epsilon^{-(p+1)})$

A BCD-ML algorithm: convergence theory

Theorem (Gratton, Mercier, R., Toint, 2023)

If f has L-Lipschitz continuos gradient and step-size $\alpha_k = \alpha < 1/L$

- **Deterministic**

$$\|\nabla f(x^{(K)})\| \leq \epsilon \to K = O\left(\frac{1}{\epsilon^2 p}\right)$$

- **Stochastic**

$$\mathbb{E}\left(\sum_{k=1}^{K} \|\nabla f(x^{(k)})\|^2\right) \leq C_1(\sigma^2) + O\left(\frac{1}{K}\right) - C_2(\sigma^2) p$$

Outline

Origins of multilevel methods: multigrid (MG)

A block coordinate descent perspective on MG methods

Application domains of the ML framework

- Physics informed neural networks (distributed)
- Image restoration (hierarchical)

Physics informed neural networks

Approximate the solution of a PDE by a neural network

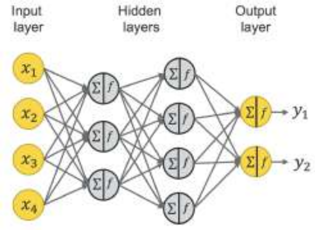
M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, 2019.

Why this approach ?

- Natural approach for nonlinear equations
- Provides analytic and continuously differentiable expression of the approximate solution
- The solution is meshless, well suited for problems with complex geometries
- The training is highly parallelizable on GPU
- Allows to alleviate the effect of the curse of dimensionality

General neural network strategy for learning problems

Neural network



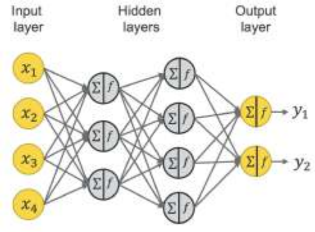
Training data

Training problem:

$$\min_{\theta \in \Theta} L(\theta) = \frac{1}{m} \sum_{i=1}^{m} (NN(\theta, x_i) - y_i)^2$$

General neural network strategy for learning problems

Neural network



Training data

Training problem:

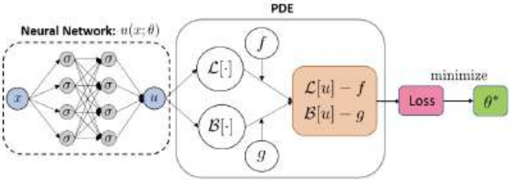
$$\min_{\theta \in \Theta} L(\theta) = \frac{1}{m} \sum_{i=1}^{m} (NN(\theta, x_i) - y_i)^2$$

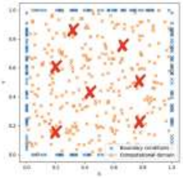
How to integrate the physical knowledge in the model?

Physics Informed Neural Networks (PINNs)

Neural network

Training data





Training problem: $\underset{\theta\in\Theta}{min} L(\theta) = L_{OBS}(\theta) + L_{PDE}(\theta)$

$$L_{OBS}(\theta) = \frac{1}{m_1} \sum_{x_i\in\Omega\cup\partial\Omega} (NN(\theta,x_i) - y_i)^2,$$

$$L_{PDE}(\theta) = \frac{1}{m_{2,i}} \sum_{x_i\in\Omega} (\mathcal{L}(NN(\theta,x_i)) - f(x_i))^2) + \frac{1}{m_{2,b}} \sum_{x_i\in\partial\Omega} (\mathcal{B}(NN(\theta,x_i)) - g(x_i))$$

How to fit the training of PINNs in the ML framework?

An important ingredient: the F-principle

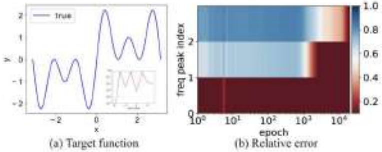
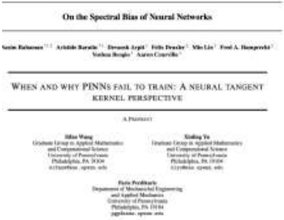


Figure 1: 1d input. (a) $f(x)$. Inset : $|\hat{f}(k)|$. (b) $\Delta_F(k)$ of three important frequencies (indicated by black dots in the inset of (a)) against different training epochs. The parameters of the DNN is initialized by a Gaussian distribution with mean 0 and standard deviation 0.1. We use a tanh-DNN with widths 1-8000-1 with full batch training. The learning rate is 0.0002. The DNN is trained by Adam optimizer [20] with the MSE loss function.



On the Spectral Bias of Neural Networks

When and why PINNs fail to train: A neural tangent kernel perspective

$\Rightarrow$ PINNs are not effective in approximating highly oscillatory solutions

How to transpose the ingredients of success of MG

Basic idea of MG: Exploiting "complementarity" between problems involved

Classical MG vs Neural networks

- Consider a minimization method and a class of problems for which this method is efficient

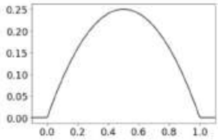
Classical MG	Neural networks
smoothing (GS orJ)	first-order (GD, SGD)
high-frequency	low-frequency

- Split the problem depending of its frequency content
- Shift the frequencies

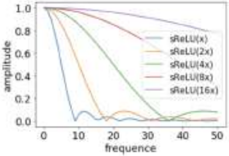
Classical MG	Neural networks
Coarser discretizations	Specialized architectures (Mscale networks)

Specialized architectures

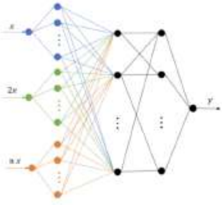
- **Mscale networks**: [Liu, Cai and Xu, (2020)] frequency-selective subnetworks + wavelet-inspired and frequency-located activation functions



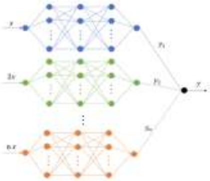
(b) sReLU



(a) sReLU



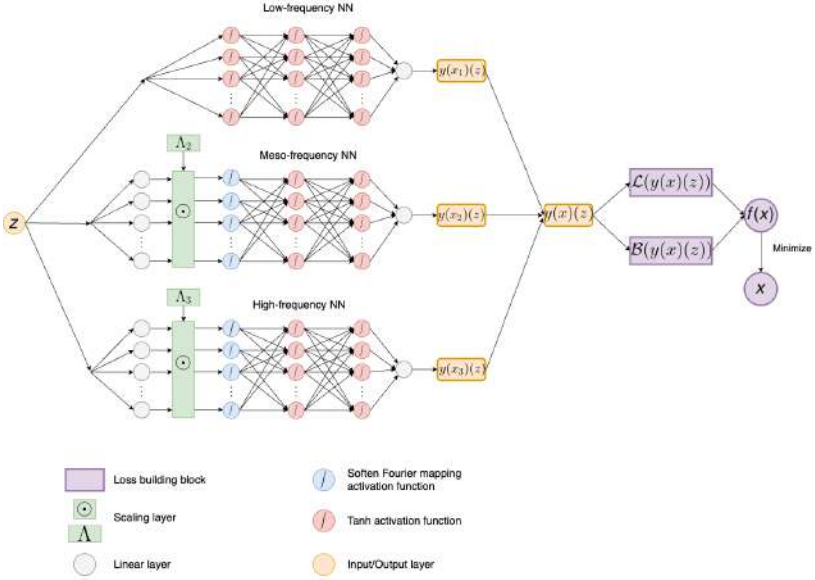
(a) MscaleDNN-1



(b) MscaleDNN-2

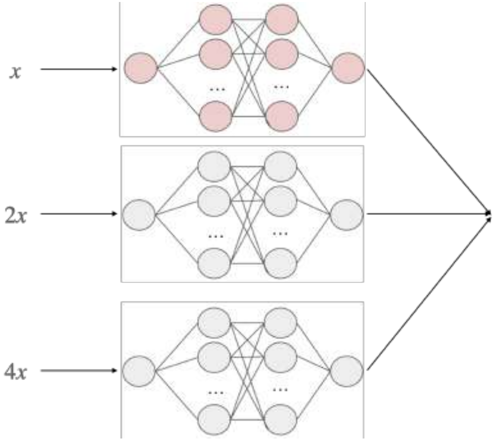
ENS DE LYON

Our architecture



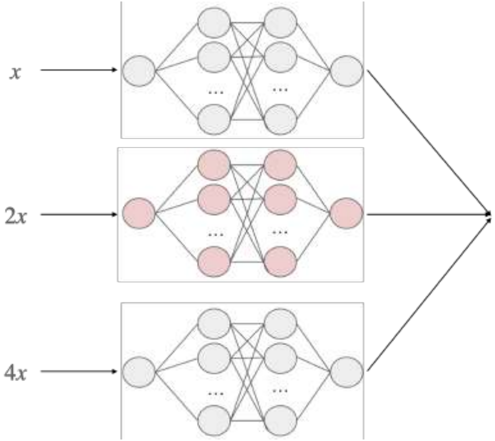
Multilevel PINNs: the training

From simultaneous training to BCD training!



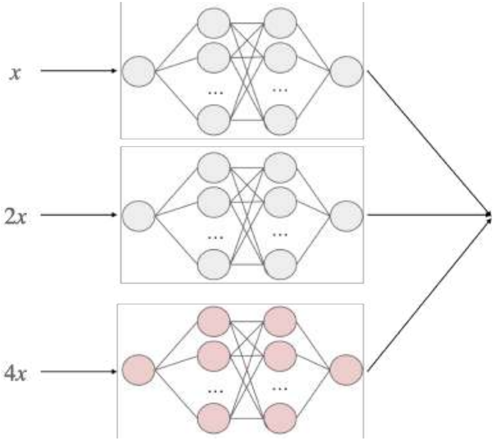
Multilevel PINNs: the training

From simultaneous training to BCD training!



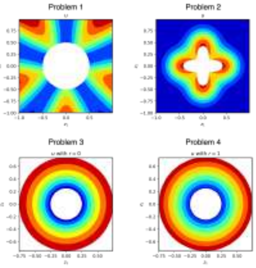
Multilevel PINNs: the training

From simultaneous training to BCD training!

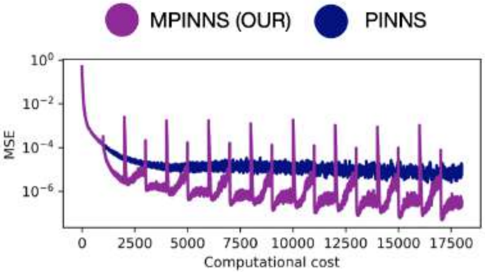


Numerical results: MSE vs iterations

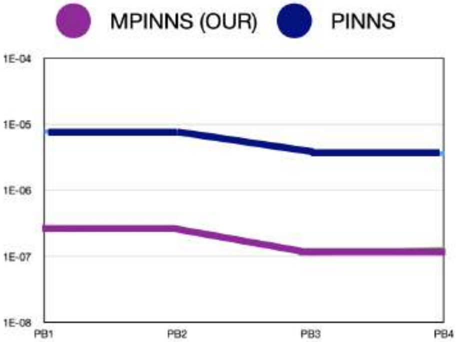
Problem: $\Delta u = f$ on Ω



Ω

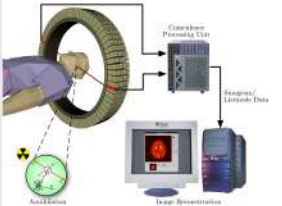


Numerical results: final MSE on average (10 runs)

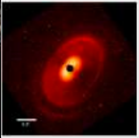


Inverse problems in imaging: various applications

Medical imaging



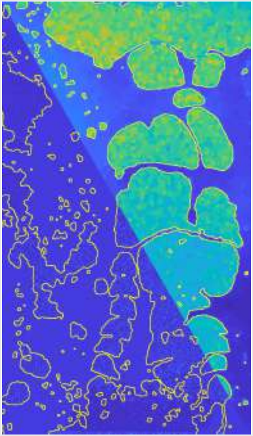
Astronomy



@ L. Denneulin

SPHERE/IRDIS

Physics



@ B. Pascal

Thanks to Nelly and Guillaume for the pictures !

Problem formulation

$$\widehat{x} \in \arg\min_{x} \frac{1}{2}\|Ax - z\|_2^2 + \lambda\|Lx\|_\star$$

with $\|Lx\|_\star$ usually sparsity inducing norm.

$\overline{x}$

z

$\widehat{x}$

Problem formulation

More generally

$$\min_x \; f(x) + g(x)$$

- f differentiable with Lipschitz gradient
- g possibly non-smooth but proximable

Classical solution methods:

- require prox computation (usually not available in closed form)
- suitable for problems of reasonable size

ML to leverage large dimensions?

Multilevel methods in nonlinear optimization (NO)

ML approaches for nonlinear smooth problems

- S.G. Nash, MG/Opt (2000)
- S. Gratton, A. Sartenaer, and P. Toint, RMTR (2008)

Multilevel methods for imaging problems?

ML approaches on smoothed image problems

- A. Javaherian and S. Holman, (tomography, 2017)
- S. W. Fung and Z. Wendy, (phase retrieval, 2020)
- J. Plier, F. Savarino, M. Kocvara, and S. Petra, (tomography, 2021)

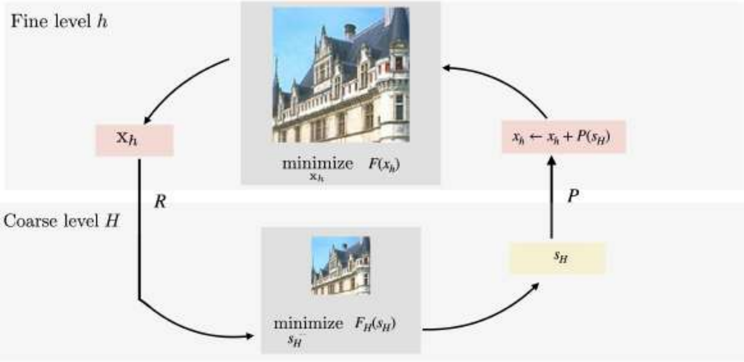
Multilevel methods for imaging problems?

ML approaches on smoothed image problems

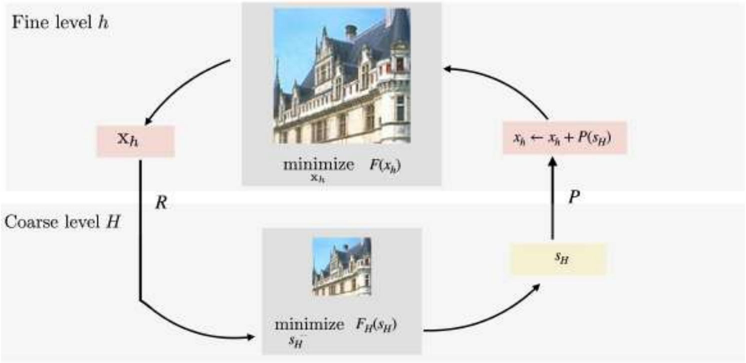
- A. Javaherian and S. Holman, (tomography, 2017)
- S. W. Fung and Z. Wendy, (phase retrieval, 2020)
- J. Plier, F. Savarino, M. Kocvara, and S. Petra, (tomography, 2021)

Extension of ML to a non-smooth setting?

An iteration of a multilevel procedure



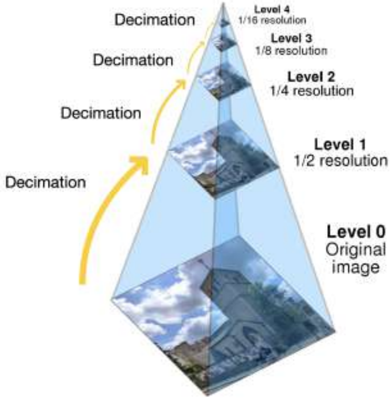
An iteration of a multilevel procedure

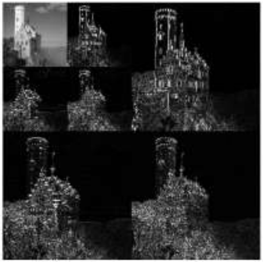


R, P?

F_H?

A hierarchy of images: R, P





Coarse model definition F_H

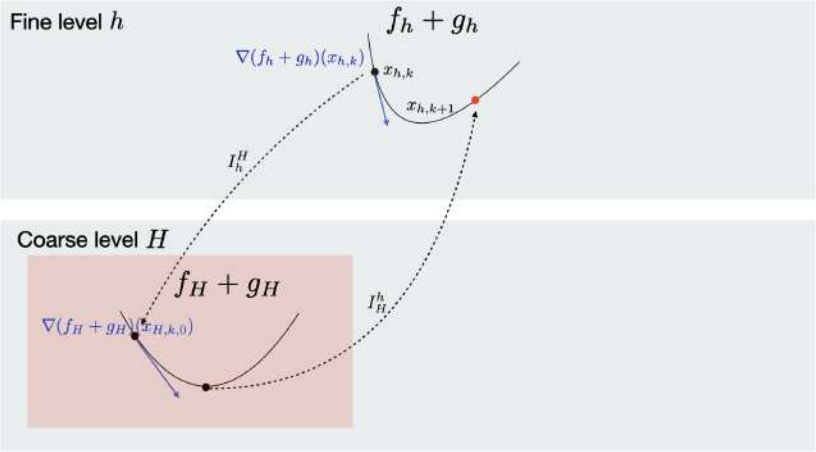
$$F(x) = \frac{1}{2}\|Ax - z\|_2^2 + \lambda\|Lx\|_1$$
$$F_H(x) = \frac{1}{2}\|A_H x_H - z\|_2^2 + \lambda\|L_H x_H\|_1$$

Coarse model definition F_H

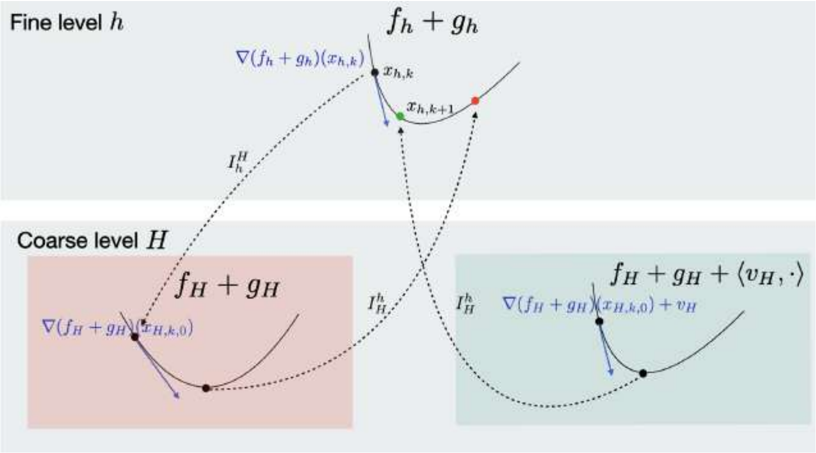
$$F(x) = \frac{1}{2}\|Ax - z\|_2^2 + \lambda\|Lx\|_1$$
$$F_H(x) = \frac{1}{2}\|A_H x_H - z\|_2^2 + \lambda\|L_H x_H\|_1$$

Is this model useful in minimizing F?

Design of F_H in smooth context: First order coherence



Design of F_H in smooth context: First order coherence



Coarse model definition F_H

$$F(x) = \frac{1}{2}\|Ax - z\|_2^2 + \lambda\|Lx\|_1$$

$$F_H(x_H) = \frac{1}{2}\|A_H x_H - z\|_2^2 + \lambda\|L_H x_H\|_1 + < v_H, x_H >$$

$$v_H = R\nabla F(x) - \nabla F_H(Rx)$$

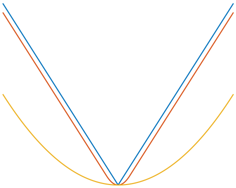
Coarse model definition F_H

$$F(x) = \frac{1}{2}\|Ax - z\|_2^2 + \lambda\|Lx\|_1$$

$$F_H(x_H) = \frac{1}{2}\|A_H x_H - z\|_2^2 + \lambda\|L_H x_H\|_1 + < v_H, x_H >$$

$$v_H = R\nabla F(x) - \nabla F_H(Rx)$$

[Parpas 2017] Nonsmooth case $\rightarrow$ smoothing!



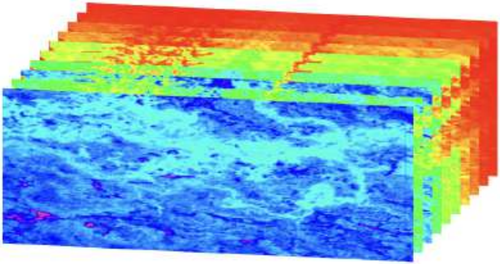
Our contributions

- Specify the method for the context of image restoration: $g(x) = \varphi(Lx)$
- Inexact proximal steps to handle state-of-the-art regularization: TV, NLTV

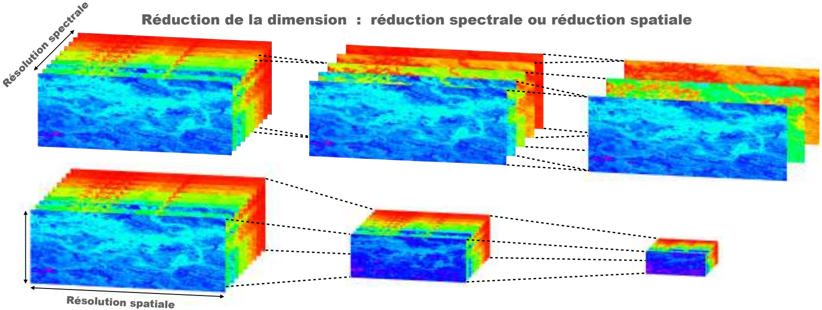
$$x_{k+1} = \tilde{\text{prox}}_{\tau\varphi\circ\mathrm{L}}\left(\bar{y}_k - \tau\nabla f(\bar{y}_k)\right)$$
$$y_{k+1} = x_{k+1} + \alpha_k(x_{k+1} - x_k)$$

  - FISTA: $\bar{y}_k = y_k$
  - IML FISTA: $\bar{y}_k = ML(y_k) \longleftrightarrow \min F_H$
- Obtain state-of-the-art convergence guarantees
- Explore definition of R, P, F_H in different contexts

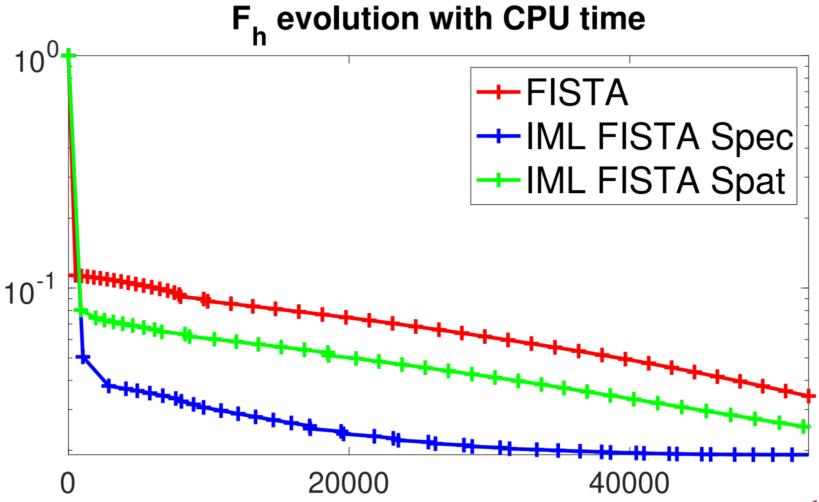
Hyperspectral images



How to build the coarse approximations?

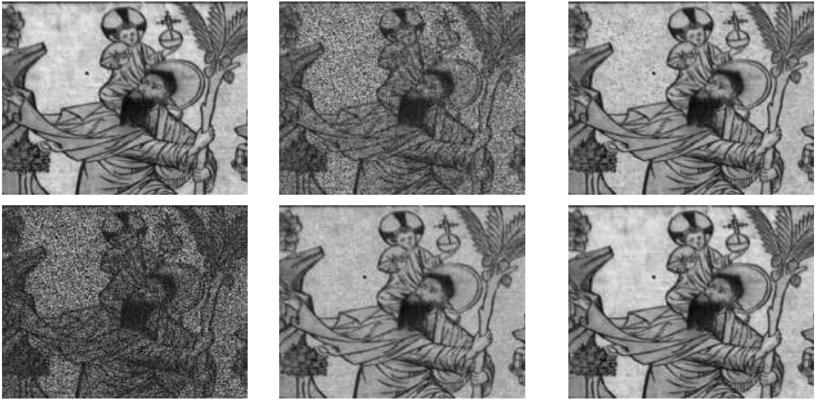


Objective function evolution



Results with Spectral IML FISTA

$\bar{x}$ (SNR) $\quad x_{2,\mathrm{FISTA}}$ (7 dB) $\quad x_{\mathrm{end,FISTA}}$ (21 dB)



z (3 dB) $\quad \mathrm{x}_{2,\mathrm{IML\ FISTA}}$ (19 dB) $\quad \mathrm{x}_{\mathrm{end,IML\ FISTA}}$ (35 dB)

UNIVERSITÉ DE LYON

ENS DE LYON

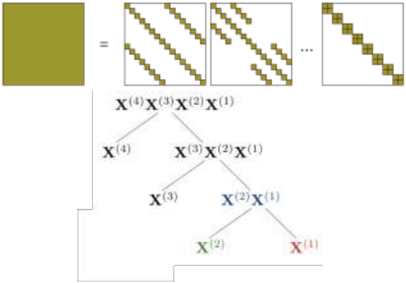
Conclusions

- We have presented a new BCD perspective on multilevel methods with unifying convergence analysis
- We have adapted the framework to two practical problems:
  - PINNs training
  - Image restoration
- We have demonstrated that exploiting multiple scales provides significant computational benefits (faster convergence).

UNIVERSITÉ DE LYON
ENS DE LYON

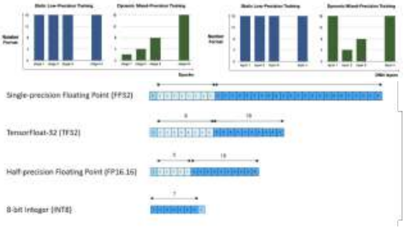
Study structure exploiting hierarchical techniques in other contexts

Sparse matrix factorization



- application to neural networks ?
- quantization?

Mixed precision training



...ongoing work with

Thank you for your attention!

A few references

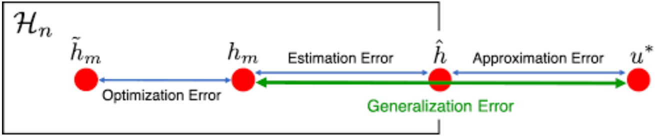
- S. Gratton, A. Sartenaer, Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization, SIAM J. Opt., 19:414–444, 2008
- S. G. Nash. A multigrid approach to discretized optimization problems. Optimization Methods and Software, 14:99–116, 2000
- W. Briggs, V. Henson, S. McCormick. A Multigrid Tutorial, SIAM, 2000
- H. Calandra, S. Gratton, E. Riccietti, X. Vasseur. On high-order multilevel optimization strategies. SIAM J. Opt., 31.1: 307-330, 2021.
- S. Wang, H. Wang, P. Perdikaris. On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks. CMAME, 384, 2021
- Z. Liu, W. Cai, Z. Xu. Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains. arXiv:2007.11207, 2020
- G. Lauga, E. Riccietti, N. Pustelnik, P. Gonçalves, IML FISTA: A Multilevel Framework for Inexact and Inertial Forward-Backward. Application to Image Restoration, preprint, 2023.
- S. Gratton, V. Mercier, E. Riccietti, Ph.L. Toint, A Block-Coordinate Approach of Multi-level Optimization with an Application to Physics-Informed Neural Networks, arXiv preprint, 2023.

UNIVERSITÉ DE LYON

ENS DE LYON

Convergence theory [Shin, Darbon, Karniadakis, 2020]

- the universality property of NN (approximation error)
- statistical sampling,
- ability of numerical optimizers (ADAM,SGD,...) to reach an approximate global optimum of nonconvex function



- $\tilde{h}_m$ our network,
- h_m a perfectly trained network on the dataset,
- $\hat{h}$ function minimizing the problem with infinitely many data,
- u^* the solution of the underlying PDE

UNIVERSITÉ DE LYON
ENS DE LYON

Convergence theory [Shin, Darbon, Karniadakis, 2020]

- L_{PINN} expected loss
- L_m empirical loss over m samples
- α Holder constant
- d dimension
- HP: the derivation is based on the probabilistic space filling arguments, assume that training data distributions cover the interior and the boundary

With high probability

$$L_{PINN}(h) \leqslant L_m(h) + C(m^{\alpha/d})$$

and

$$L_{PINN}(h_m) \leqslant C(m^{\alpha/d})$$

with $h_m \in H_n$ minimizer of L_m If PDE is linear (elliptic or parabolic)

$$\lim_{m\to\infty} h_m = u^*$$

(seq of minimizers conv uniformly to PDE sol in infinite data regime)

Inexact FISTA

$$\min_x f(x) + \varphi(\mathrm{L}x)$$

[Aujol, Dossal, 2015]:

$$\begin{aligned} x_{k+1} &\approx_{\epsilon_k} \operatorname{prox}_{\tau\varphi\circ\mathrm{L}}\left(y_k - \tau\nabla f(y_k) + e_k\right) \\ y_{k+1} &= x_{k+1} + \alpha_k(x_{k+1} - x_k) \end{aligned}$$

Contribution: update y_k through a multilevel step.

IML FISTA

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \varphi(\mathrm{L}x)$$

$$x_{k+1} = \tilde{\text{prox}}_{\tau\varphi\circ\mathrm{L}}\left(\bar{y}_k - \tau\nabla f(\bar{y}_k)\right)$$
$$y_{k+1} = x_{k+1} + \alpha_k(x_{k+1} - x_k)$$

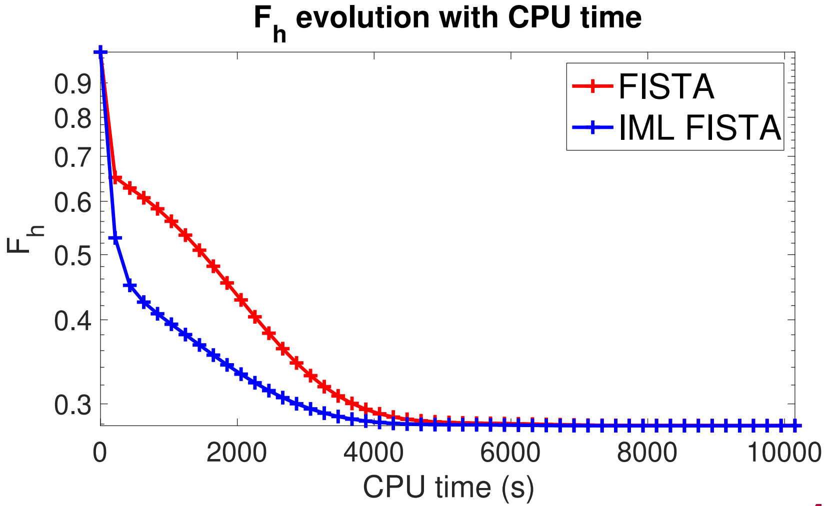
- FISTA: $\bar{y}_k = y_k$
- IML FISTA: $\bar{y}_k = ML(y_k) \longleftrightarrow \min F_H$

$$F_H(x) := \frac{1}{2}x^T(P^TAP)x - b^TPx + {}^{\gamma}\varphi_H \circ L_H + {v_H}^T x$$

$${}^{\gamma}\varphi, {}^{\gamma}\varphi_H \to v_H$$

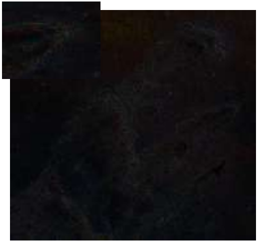
Image reconstruction with NLTV prior

Evolution of F_h for a $N_h = 2048 \times 2048 \times 3$ image

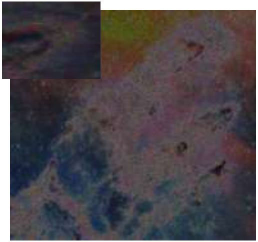


Reconstruction after 2 iterations with NLTV

z



x_2 FISTA



x_2 IML FISTA

The hierarchical context

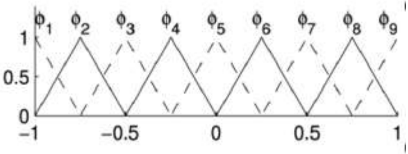
Example: classical MG

$$f(x) = \frac{1}{2}x^T A x + x^T b$$

$$y_1(x_1) = \sum_{j=1}^{n_1} x_{1,j} b_j$$

$$y_2(x_2) = \sum_{j=1}^{n_1} (Px_2)_j \, b_j$$

$$\mathcal{F} = \mathrm{span}\{b_j\}_{j=1}^{m}, \; m = n_1$$



In this context, we have that

$$\mathcal{A}_2 = \left\{ \sum_{j=1}^{m} (Px_2)_j \, b_j \right\} \subset \left\{ \sum_{j=1}^{m} (x_{1,j} + (Px_2)_j) \, b_j \right\} = \mathcal{A}_1 = \mathcal{A}_{12}.$$