

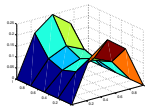
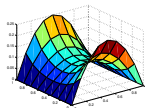
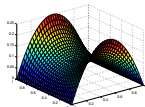
A multilevel stochastic method with application to finite sum minimization

Elisa Riccietti - LIP-ENS Lyon

Filippo Marini - UNIBO

Margherita Porcelli - UNIFI

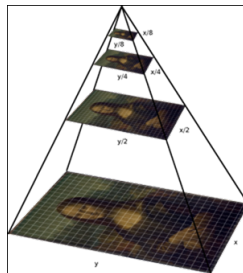
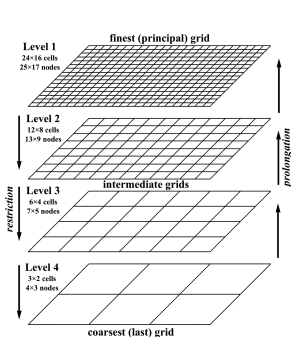
SIOPT 2026 - Edinburgh



The concept

AIM? Given a large scale optimization problem, reduce the computational cost of the solution process

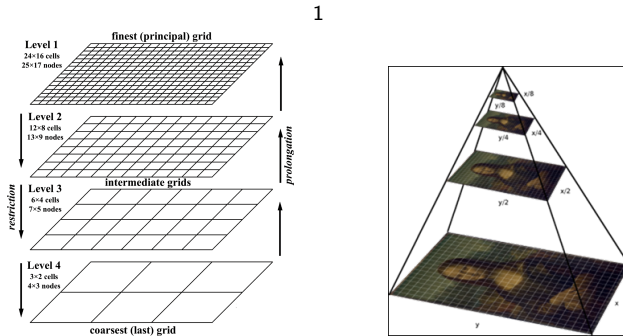
HOW? Exploit the **structure** to work at multiple resolutions



The concept

AIM? Given a large scale optimization problem, reduce the computational cost of the solution process

HOW? Exploit the **structure** to work at multiple resolutions



Multilevel methods (ML)

Outline

- ▶ Part I: Origins of multilevel methods: multigrid
- ▶ Part II: Multilevel methods for nonlinear optimization
- ▶ Part III: Multilevel stochastic methods

Part I

Origins of multilevel methods: multigrid

Numerical solution of linear PDEs

Given $f : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$, $g : \partial\Omega \rightarrow \mathbb{R}$ and a linear differential operator D , find $u : \mathbb{R}^d \rightarrow \mathbb{R}$ that solves

$$D(u(x)) = f(x) \text{ in } \Omega$$

$$u(x) = g(x) \text{ in } \partial\Omega$$

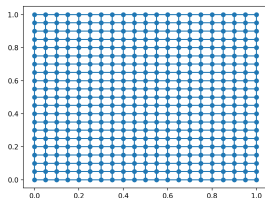


Discretization



Linear system

$$Au = f$$



Example: $d=2$, $\Omega = [0, 1] \times [0, 1]$

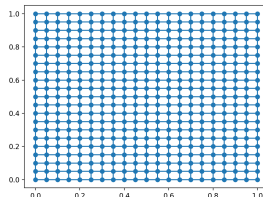
The size of the grid impacts:

- ▶ the **size** of the discrete problem
- ▶ the **accuracy** of the solution approximation

Multigrid methods

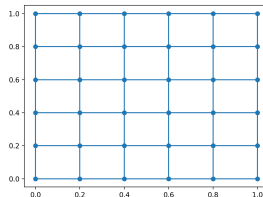
Aim: accelerate fixed point methods by exploiting multiple grids

► Fine grid: $A^h u^h = f^h$

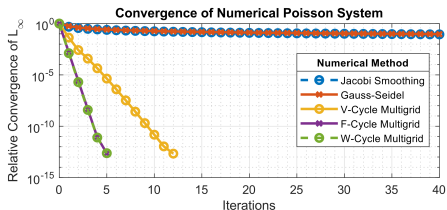
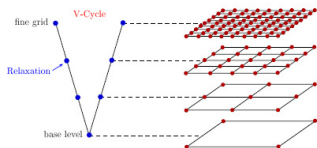


$R \Downarrow$ $P \Uparrow$

► Coarse grid: $A^H u^H = f^H$



General multigrid methods - V-cycle



Images taken from https://en.wikipedia.org/wiki/Multigrid_method

Part II

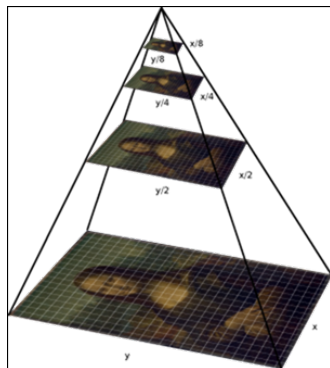
From linear to nonlinear: multilevel
optimization methods

From linear to nonlinear: multilevel methods

Multigrid methods work for linear systems: $Ax = b$

But what happens if we want to tackle a more general **optimization problem** with a hierarchical structure ?

$$\min_x f(x)$$



Classical iterative optimization methods

- Build a **model** ($B_k \sim \nabla^2 f(x_k)$ or $B_k = 0$):

$$f(x_k + s) \simeq T_k(s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T B_k s$$

Classical iterative optimization methods

- ▶ Build a **model** ($B_k \sim \nabla^2 f(x_k)$ or $B_k = 0$):

$$f(x_k + s) \simeq T_k(s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T B_k s$$

- ▶ Compute a **step** s_k ($r(\lambda_k)$ regularization term):

$$\min_s m_k(s) = T_k(s) + r(\lambda_k), \quad \lambda_k > 0$$

bottleneck: subproblem solution

Possible solution: multilevel methods



MG/OPT: *A multigrid approach to discretized optimization problems*, Nash, 2000



RMTR: *Recursive trust-region methods*, S. Gratton, A. Sartenaer and Ph. L. Toint, 2008

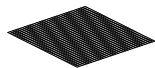
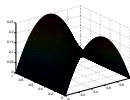
Hierarchy of problems ($n_{\ell_{\max}} > n_{\ell_{\max}-1} > \dots > n_0$)



Finest Level $f^{\ell_{\max}} : \mathbb{R}^{n_{\ell_{\max}}} \rightarrow \mathbb{R}$

Restriction $\downarrow R^{\ell_{\max}}$

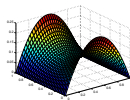
$P^{\ell_{\max}}$ $\uparrow$ Prolongation



Fine Level $f^{\ell_{\max}-1} : \mathbb{R}^{n_{\ell_{\max}-1}} \rightarrow \mathbb{R}$

Restriction $\downarrow R^{\ell_{\max}-1}$

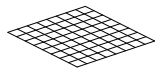
$P^{\ell_{\max}-1}$ $\uparrow$ Prolongation



$\vdots$

Restriction $\downarrow R^2$

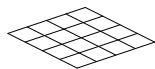
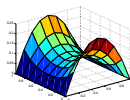
P^2 $\uparrow$ Prolongation



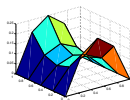
Coarse Level $f^2 : \mathbb{R}^{n_2} \rightarrow \mathbb{R}$

Restriction $\downarrow R^1$

P^1 $\uparrow$ Prolongation



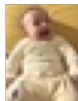
Coarsest Level $f^0 : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$



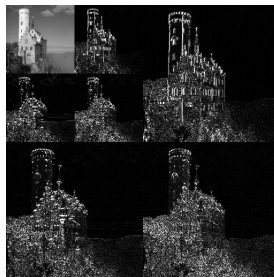
Example 1: image restoration



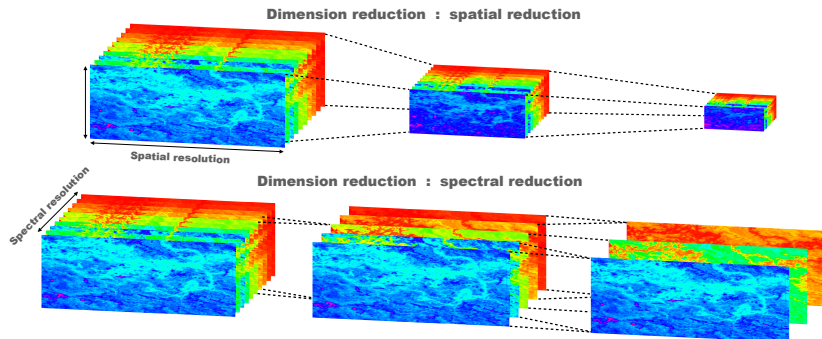
Decimation



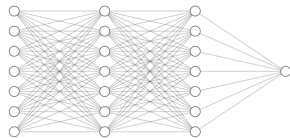
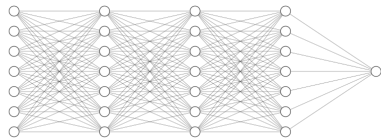
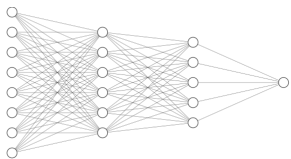
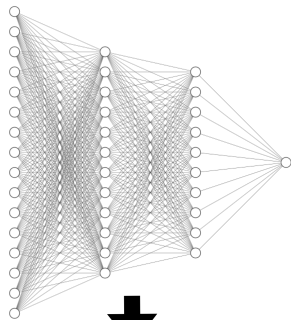
Decimation



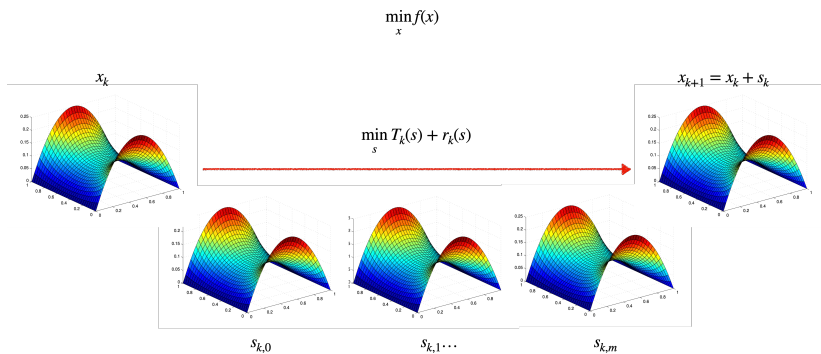
Example 2: hyperspectral imaging



Example 3: neural networks



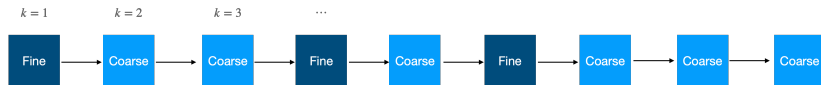
Classical one-level method



Multilevel strategy: step computation

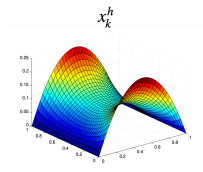
Two choices:

1. Classical fine step
2. Coarse step

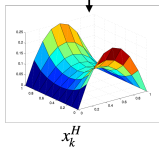


Coarse step

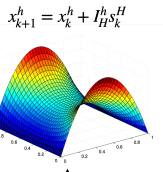
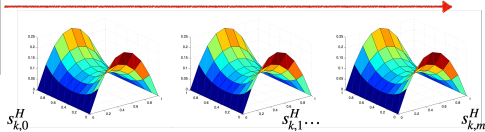
$$\min_{x^h} f^h(x^h) \sim \min_{x^H} f^H(x^H)$$



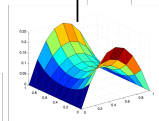
R



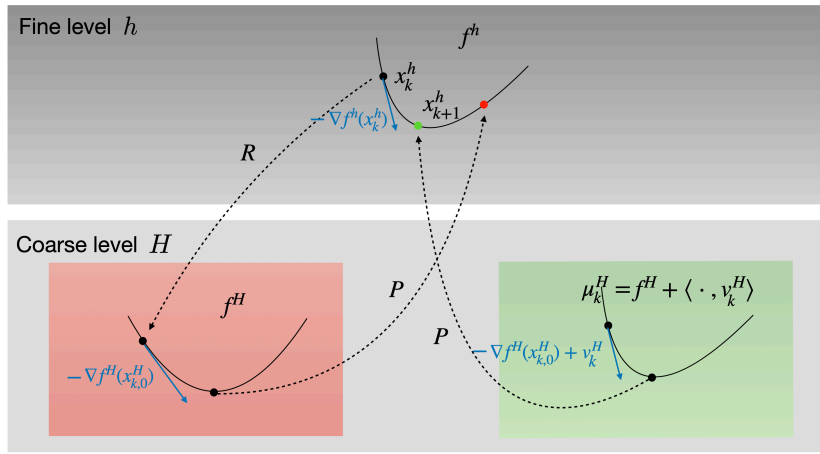
$$\min_{s^H} \mu_k^H(s^H)$$



P



Definition of lower level model: first order coherence

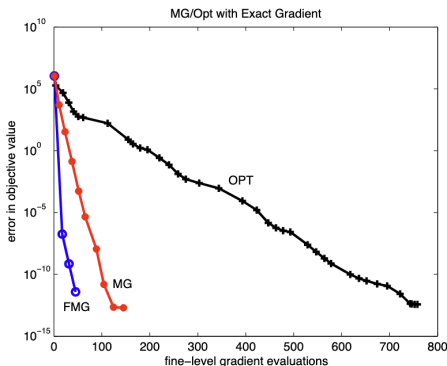


$$v_k^H = R \nabla f^h(x_k^h) - \nabla f^H(x_k^H) \rightarrow \underbrace{\nabla \mu_k^H(0^H) = R \nabla f^h(x_k^h)}_{\text{first-order coherence}}$$

Numerical results for 5 levels MG/OPT with line-search

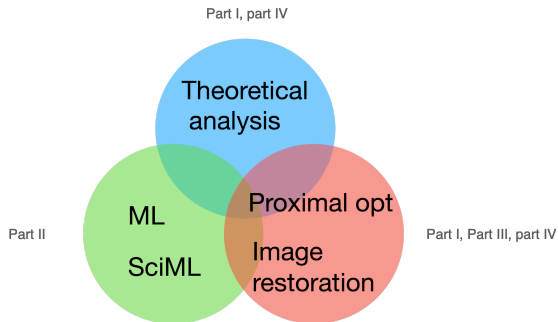
Problem:

$$\min_a F(a) = \frac{1}{2} \int_0^\pi \left(\frac{\partial u}{\partial x_2}(x_1, 0) - \phi(x_1) \right)^2 dx_1,$$



Using Inexact Gradients in a Multilevel Optimization Algorithm, M. Lewis, S. G. Nash, 2013

A very active field



Part III

Stochastic multilevel optimization

Stochastic optimization problems

What if my problem is stochastic?

$$\min_x f(x)$$

where f can only be computed **with some noise**.

Example - *Expected risk minimization problems*

Given a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, a parametric model $m_x : \mathcal{Z} \rightarrow \mathcal{Y}$ and a probability distribution P on $\mathcal{Z} \times \mathcal{Y}$

$$\min_{x \in \mathbb{R}^n} \mathbb{E}_P[\mathcal{L}(y, m_x(z))]$$

Example - *Montecarlo simulations*

Let ξ be a random variable defined on the probability space (Ω, P) .

Let $f(x) = \mathbb{E}(F(x, \xi))$, with $F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$.

Limitation of existing methods

Stochastic methods:

- ▶ Accuracy of the function approximations **grows** asymptotically, which leads to iterations that are more and more expensive:

$$\|f(x) - \tilde{f}(x)\| \leq \epsilon_k, \quad \epsilon_k \rightarrow 0$$

- ▶ Such methods control the accuracy of the function estimates but not the **size of the variables**.

Multilevel framework:

- ▶ Limited to **deterministic** contexts
- ▶ Still needs to handle **full problem** at the finest scale

Limitation of existing methods

Stochastic methods:

- ▶ Accuracy of the function approximations **grows** asymptotically, which leads to iterations that are more and more expensive:

$$\|f(x) - \tilde{f}(x)\| \leq \epsilon_k, \quad \epsilon_k \rightarrow 0$$

- ▶ Such methods control the accuracy of the function estimates but not the **size of the variables**.

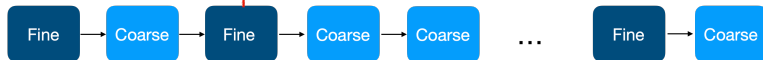
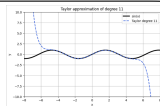
Multilevel framework:

- ▶ Limited to **deterministic** contexts
- ▶ Still needs to handle **full problem** at the finest scale

→ **Stochastic multilevel methods!**

Stochastic multilevel iterations

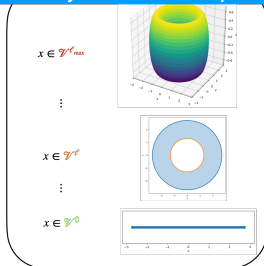
$$\|f(x) - f^{\text{fine}}(x)\| \leq \epsilon_k$$



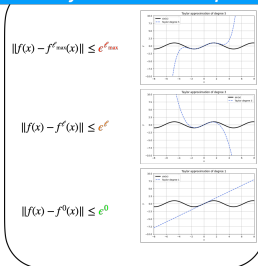
$$\dim(\mathcal{V}^0) \leq \dots \leq \dim(\mathcal{V}^\ell) \leq \dots \leq \dim(\mathcal{V}^{\ell_{\max}})$$

$$\epsilon^0 \geq \dots \geq \epsilon^\ell \geq \dots \geq \epsilon^{\ell_{\max}}$$

Hierarchy in the variables space



Hierarchy in the function space



Example - Expected risk minimization

- Fine level:

$$\tilde{f}_k(x) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} f^{(i)}(x) \quad x \in \mathbb{R}^n, \quad |\mathcal{S}_k| \nearrow$$

- Coarse level:

Hierarchy in the variables space:

Randomly draw

$$\mathcal{V}^0 \subseteq \dots \subseteq \mathcal{V}^\ell \subseteq \dots \subseteq \mathcal{V}^{\ell_{\max}} \subseteq \mathbb{R}^n$$

$$\phi_k^\ell(x^\ell) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} f^{(i)}(x^\ell)$$

$$x^\ell \in \mathcal{V}^\ell$$

Hierarchy in the function space:

Randomly draw

$$\mathcal{S}_k^0 \subseteq \dots \subseteq \mathcal{S}_k^\ell \subseteq \dots \subseteq \mathcal{S}_k^{\ell_{\max}} \subseteq \mathcal{S}_k$$

$$\phi_k^\ell(x) = \frac{1}{|\mathcal{S}_k^\ell|} \sum_{i \in \mathcal{S}_k^\ell} f^{(i)}(x)$$

$$x \in \mathbb{R}^n$$

Proposed stochastic multilevel method (2 levels)

Mu^ℓSTREG: First order **adaptive regularization method**

$$m_k(x_k + s) = \varphi_k(s) + \frac{\lambda_k \|g_k\|}{2} \|s\|^2$$

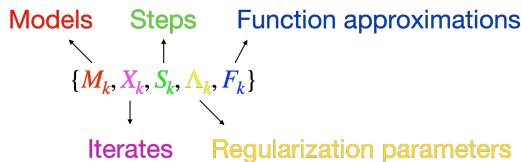
► Fine level:

$$\varphi_k(s) = f_k + g_k^T s \quad f_k \approx f(x_k), \quad g_k \approx \nabla f(x_k)$$

► Coarse level:

$$\varphi_k(s) = \phi_k(Rx_k + s) + \underbrace{(Rg_k - \nabla \phi_k(Rx_k))^T s}_{\text{first-order coherence}}$$

The theory - what are we interested in?



The stochastic process $\{M_k, X_k, S_k, \Lambda_k, F_k\}$, under certain conditions on the sequences $\{M_k\}$ and $\{F_k\}$ has desirable convergence properties with **probability one**.

The theory - assumptions [simplified]

- ▶ $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\phi_k : \mathbb{R}^{n_c} \rightarrow \mathbb{R}$ with $n \geq n_c$ have Lipschitz continuous gradients
- ▶ The model M_k to be κ -fully linear around X_k with probability α (only at fine level!)
- ▶ The function approximations F_k to be sufficiently accurate (ϵ_f) with probability β

The theory - results

Under suitable choices for ϵ_f , α and β can be chosen so that, if the assumptions hold for these values, then the sequence of regularization parameters $\{\Lambda_k\}$ satisfies

$$\sum_{k=0}^{\infty} \frac{1}{\Lambda_k^2} < \infty$$

and the sequence of iterates $\{X_k\}$ satisfy

$$\lim_{k \rightarrow \infty} \|\nabla_x f(X_k)\| = 0$$

almost surely.

Numerical results - Finite sum minimization

The problem

$$\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{2N} \sum_{i=1}^N \left(y_i - \frac{1}{1 + \exp(-y_i x^T z_i)} \right)^2,$$

with $(z_i, y_i) \in \mathbb{R}^n \times \{0, 1\}$, for every $i = 1, \dots, N$.

The methods

- ▶ MU¹STREG: one level ($\approx$ STORM [Chen, et al. 2018])
- ▶ MU³STREG: three levels
- ▶ Adagrad [J. Duchi, E. Hazan, and Y. Singer. (2011)]
- ▶ SVRG [R. Johnson and T. Zhang. (2013)]

$$x_{k+1} = x_k - \alpha \left[\frac{1}{b} \sum_{i_k \in I_k} \nabla f^{i_k}(x_k) + \left(\nabla f(\tilde{x}_0) - \frac{1}{b} \sum_{i_k \in I_k} \nabla f^{i_k}(\tilde{x}_0) \right) \right],$$

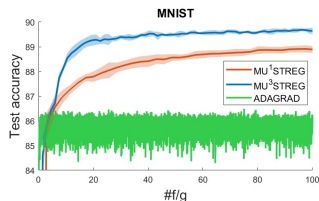
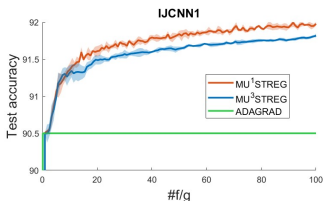
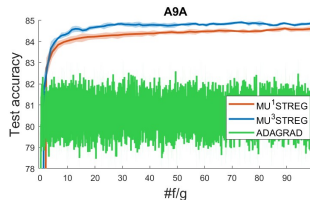
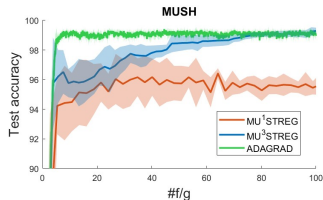
where $\tilde{x}_0$ is updated every m iterations

Results: comparison against MU¹STREG and SVRG

MUSH			
	SVRG	MU ¹ STREG	MU ³ STREG
Avg. %tA	98.69	97.68	97.74
StD %tA	0.25	0.50	0.48
Avg. #f/g	341.89	216.78	35.48
StD #f/g	722.65	30.69	9.95
A9A			
	SVRG	MU ¹ STREG	MU ³ STREG
Avg. %tA	85.00	84.65	84.83
StD %tA	0.05	0.04	0.07
Avg. #f/g	840.77	207.68	90.87
StD #f/g	300.51	53.97	30.31
IJCNN1			
	SVRG	MU ¹ STREG	MU ³ STREG
Avg. %tA	91.68	91.10	91.13
StD %tA	0.00	0.25	0.09
Avg. #f/g	553.87	6.20	7.09
StD #f/g	1.64	0.97	0.64
MNIST			
	SVRG	MU ¹ STREG	MU ³ STREG
Avg. %tA	90.43	89.80	89.82
StD %tA	0.13	0.04	0.05
Avg. #f/g	17843.40	2923.02	414.15
StD #f/g	6336.56	497.15	16.37

Table: Comparison between MU¹STREG, MU³STREG and SVRG. Average of maximum classification accuracy reached and number of evaluations, with corresponding standard deviation.

Results: comparison against MU¹STREG and Adagrad



Classification accuracy on the testing set along the number of weighted evaluations of gradients and functions.

Conclusions

Multilevel framework:

- ▶ Flexible framework to accelerate large scale problems, adaptable to different contexts

MU^ℓSTREG

- ▶ allows for hierarchies in the both variable and function spaces;
- ▶ is the first stochastic framework for multilevel methods;
- ▶ Applied to **finite sum problems**: outperforms tested methods; wrt to the classical deterministic ML framework: **it does not need the evaluation of the function/gradient over the full samples set along all the iterations.**

F. Marini, M. Porcelli, E. Riccietti, **A multilevel stochastic regularized first-order method with application to finite sum minimization**, 2024, *arXiv:2412.11630*, accepted in Mathematics of computation

STOP workshop



STructured OPTimization for inverse and learning problems

28 september - 01 october 2026

ENS Lyon

<https://perso.ens-lyon.fr/elisa.riccietti/STOP.php>

STOP workshop



STructured OPTimization for inverse and learning problems

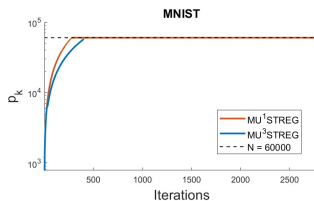
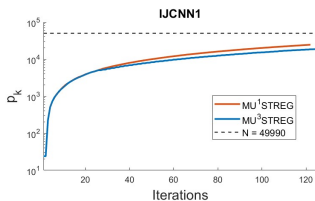
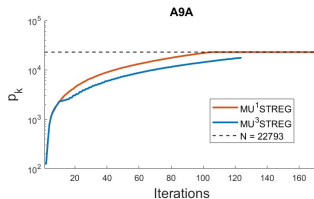
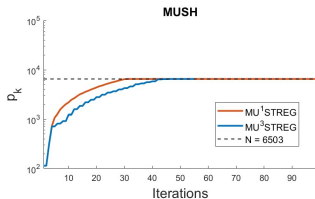
28 september - 01 october 2026

ENS Lyon

<https://perso.ens-lyon.fr/elisa.riccietti/STOP.php>

Thank you for your attention!

Numerical results - Finite sum minimization



Cardinality of the sample set at the finest level and the full size N along the iterations. $|S_k^{\ell_{\max}}| = p_k := \min\{N, \max\{100k + n + 2, \lambda_k^2\}\}$

The theory - assumptions

Let $\mathcal{F}_{k-1/2}^{MF}$ be the σ -algebra generated by $M_0, \dots, M_k$ and $F_0, \dots, F_{k-1}$ and $\mathcal{F}_{k-1}^{MF}$ be the σ -algebra generated by $M_0, \dots, M_{k-1}$ and $F_0, \dots, F_{k-1}$.

- ▶ $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\phi_k : \mathbb{R}^{n_c} \rightarrow \mathbb{R}$ with $n \geq n_c$ have Lipschitz continuous gradients
- ▶ M_k **α -probabilistically κ -fully linear models**: $\mathbb{P}(I_k | \mathcal{F}_{k-1}^{MF}) \geq \alpha$,
 $I_k = \{M_k \text{ is a } \kappa\text{-fully linear model of } f \text{ around } X_k\}$
- ▶ F_k **β -probabilistically ϵ_f -accurate estimates**:
 $\mathbb{P}(J_k | \mathcal{F}_{k-1/2}^{MF}) \geq \beta$,
 $J_k = \{F_k, F_k^s \text{ are } \epsilon_f\text{-accurate estimates of } f(x_k) \text{ and } f(x_k + s_k), \text{ respectively, around } X_k\}$

$$|f_k - f(x_k)| \leq \frac{\epsilon_f}{\lambda_k^2} \text{ and } |f_k^s - f(x_k + s_k)| \leq \frac{\epsilon_f}{\lambda_k^2}.$$

Convergence theory: definitions

Definition 1

Suppose that ∇f is Lipschitz continuous. Given $\lambda_k > 0$, a function m is a κ -fully linear model of f around the iterate x_k provided for $\kappa = (\kappa_f, \kappa_g)$, that for all y in a neighborhood of x_k :

$$\begin{aligned}\|\nabla_x f(y) - \nabla_x m(y)\| &\leq \frac{\kappa_g}{\lambda_k}, \\ |f(y) - m(y)| &\leq \frac{\kappa_f}{\lambda_k^2}.\end{aligned}$$

Definition 2

A sequence of random models $\{M_k\}$ is said to be α -probabilistically κ -fully linear with respect to the corresponding sequence $\{X_k, \Lambda_k\}$ if the events

$$I_k = \{M_k \text{ is a } \kappa \text{-fully linear model of } f \text{ around } X_k\}$$

satisfy the condition

$$\mathbb{P}(I_k | \mathcal{F}_{k-1}) \geq \alpha$$

Convergence theory: definitions

Definition 3

The estimates f_k and f_k^s are said to be ϵ_f -accurate estimates of $f(x_k)$ and $f(x_k + s_k)$ respectively, for a given λ_k if

$$|f_k - f(x_k)| \leq \frac{\epsilon_f}{\lambda_k^2} \text{ and } |f_k^s - f(x_k + s_k)| \leq \frac{\epsilon_f}{\lambda_k^2}.$$

Definition 4

A sequence of random estimates $\{F_k, F_k^s\}$ is said to be β -probabilistically ϵ_f -accurate with respect to the corresponding sequence $\{X_k, \Lambda_k, S_k\}$ if the events

$$J_k = \{F_k, F_k^s \text{ are } \epsilon_f\text{-accurate estimates of } f(x_k) \\ \text{and } f(x_k + s_k), \text{ respectively, around } X_k\}$$

satisfy the condition

$$\mathbb{P}(J_k | \mathcal{F}_{k-1/2}^{M \cdot F}) \geq \beta,$$

Data sets

The data sets are divided into a training set and a testing set as specified in Table below.

Data set	nr. of features (n)	Training set size (N)	Testing set size (N_t)
MNIST	784	60000	10000
MUSH	112	6503	1621
A9A	123	22793	9768
IJCNN1	22	49990	91701

Table: Data sets with number of features n and number of instances of the training set N and the testing set N_t .