Multi**G**rid **Prox**imal Gradient Descent

Andersen Ang

ECS, USouthampton, UK

Homepage `angms.science`

Andersen Ang, Hans De Sterck, Steve Vavasis,
"MGProx: A nonsmooth multigrid proximal gradient method with adaptive restriction for strongly convex optimization",
SIAM Journal of Optimization, to appear, 2024

arXiv 2302.04077 joint work with

Hans De Sterck

Steve Vavasis

25th International Symposium on Mathematical Programming, Montreal, Canada, July 26, 2024

Not following paper's notation

Algorithm 4.1 L-level `MGProx` with V-cycle structure for an approximate solution of (1.1)

Initialize x_0^1 and the full version of $R_{\ell\to\ell+1}, P_{\ell+1\to\ell}$ for $\ell \in \{0, 1, \dots, L-1\}$
for $k = 1, 2, \dots$ **do**
  Set $\tau_{-1\to 0}^{k+1} = 0$
  for $\ell = 0, 1, \dots, L-1$ **do**
    $y_\ell^{k+1} = \mathrm{prox}_{\frac{1}{L_\ell} g_\ell}\Big(x_\ell^k - \dfrac{\nabla f_\ell(x_\ell^k) - \tau_{\ell-1\to\ell}^{k+1}}{L_\ell}\Big)$ pre-smoothing
    $x_{\ell+1}^k = R_{\ell\to\ell+1}(y_\ell^{k+1})\, y_\ell^{k+1}$ restriction to next level
    $\tau_{\ell\to\ell+1}^{k+1} \in \underline{\partial F_{\ell+1}(x_{\ell+1}^k)} - R_{\ell\to\ell+1}(y_\ell^{k+1})\, \underline{\partial F_\ell(y_\ell^{k+1})}$ create tau vector
  end for
  $w_L^{k+1} = \operatorname*{argmin}_\xi \Big\{ F_L^\tau(\xi) \coloneqq F_L(\xi) - \langle \tau_{L-1\to L}^{k+1}, \xi\rangle \Big\}$ solve the level-L coarse problem
  for $\ell = L-1, L-2, \dots, 0$ **do**
    $z_\ell^{k+1} = y_\ell^{k+1} + \alpha P_{\ell+1\to\ell}(w_{\ell+1}^{k+1} - x_{\ell+1}^k)$ coarse correction
    $w_\ell^{k+1} = \mathrm{prox}_{\frac{1}{L_\ell} g_\ell}\Big(z_\ell^{k+1} - \dfrac{\nabla f_\ell(z_\ell^{k+1}) - \tau_{\ell-1\to\ell}^{k+1}}{L_\ell}\Big)$ post-smoothing
  end for
  $x_0^{k+1} = w_0^{k+1}$ update the fine variable
end for

ℓ	level
k	iteration counter
$\boldsymbol{x}_\ell$	variable at level ℓ before ProxGrad
$\boldsymbol{y}_\ell$	variable at level ℓ after ProxGrad
f_ℓ	smooth part at level ℓ
g_ℓ	nonsmooth part at level ℓ
L_ℓ	Lipschitz constant of ∇f_ℓ at level ℓ
L	number of levels

- Too many subscripts, hard to read
- idea is important, not the maths detail
- Notation: Big and small

Setup

$$\operatorname*{argmin}_{\boldsymbol{x}} \; f(\boldsymbol{x}) + g(\boldsymbol{x})$$

- everything in finite dimensional Euclidean space
- not about
  - $\mathbb{E}$
  - $\nabla^2 f$
  - linear
  - big N in finite sum $\frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x})$
  - application

- $f : \mathbb{R}^n \to \mathbb{R}$ is
  - μ-strongly convex
  - L-smooth
- $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is
  - convex
  - proper
  - possibly nonsmooth
    - further assume a single point of non-differentiability
  - separable: $g(\boldsymbol{x}) = g_1(x_1) + g_2(x_2) + \cdots$
    - e.g. $\max\{\cdot\}$, $\|\cdot\|_1$
  - proximable

What is the idea

- goal: solve

$$\operatorname*{argmin}_{\boldsymbol{x}_{\mathsf{Big}} \in \mathbb{R}^n} f_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}}) + g_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}})$$

- n big, expensive

  e.g. galaxy image (Lauga et. al., IML-FISTA.)

What is the idea

- goal: solve

$$\operatorname*{argmin}_{\boldsymbol{x}_{\mathsf{Big}} \in \mathbb{R}^n} f_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}}) + g_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}})$$

- n big, expensive

  e.g. galaxy image (Lauga et. al., IML-FISTA.)

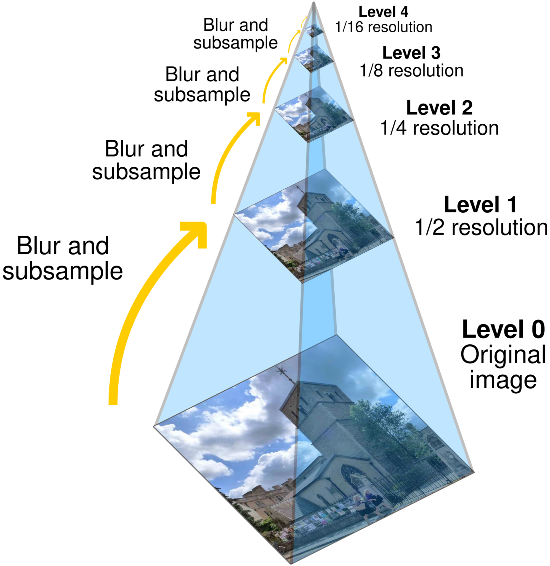
- natural idea: make use of *subspace*

$$\operatorname*{argmin}_{\boldsymbol{x}_{\mathsf{small}}} f_{\mathsf{small}}(\boldsymbol{x}_{\mathsf{small}}) + g_{\mathsf{small}}(\boldsymbol{x}_{\mathsf{small}})$$

  - how to define small problem?
  - how to create small problem?
  - how to solve small problem?

Idea is old

- Caratheodory in 1910s
- Low rank by Eckart-Young in 1936
- PDEs: multigrid in 1962[a]
- CFD: model order reduction in 1967
- Linear programming: aggregation in 1977
- Computer vision: Pyramids in 1980s
- Natural language processing: Topic modelling in 1980s
- Wavelet in 1989
- Sparse linear statistic model in 1996
- Image segmentation: SuperPixel in 2000s
- Graph: compressing network by SuperNode in 2000s
- Core-set in 2010s
- Knowledge distillation



[a] R. P. Fedorenko, "A relaxation method for solving elliptic difference equations", USSR Computational Mathematics and Mathematical Physics, 1962

What's there

- Multigrid for PDEs, **a whole field**
- mostly smooth problems
- two types
  - Full approximation scheme (FAS) / τ-approximation scheme (1st-order method)
  - Newton-multigrid method (2nd-order method)
- works for nonsmooth problem exist, but
  - use smoothing, e.g. $|x| \to \sqrt{x^2 + \epsilon}$
  - only simple constraint

What's new

- Extend the FAS to nonsmooth
- No smoothing, face subdifferential directly
- Main output: proving it works
  - (new) adaptive restriction
  - consistent optimality between two worlds
  - descents
- other fancy stuffs
  - convergence rate $1/k$
  - acceleration rate $1/k^2$
  - good-looking curves

How to go from f_{Big} to f_{small}: Subspace

$$
\begin{aligned}
f_{\text{small}} &:= f_{\text{Big}} \circ \boldsymbol{R} \\
f_{\text{small}}(\boldsymbol{x}_{\text{small}}) &:= f_{\text{Big}}(\boldsymbol{R}\boldsymbol{x}_{\text{Big}}) \\
\boldsymbol{x}_{\text{small}} &:= \boldsymbol{R}\boldsymbol{x}_{\text{Big}}
\end{aligned}
$$

- $\boldsymbol{R}$ is called *restriction* in PDEs
  - $\boldsymbol{R}$ is a short-wide matrix that maps from $\mathbb{R}^N$ to $\mathbb{R}^n$
  - if $n < N$ short-wide: not one-to-one but many-to-one

How to go from f_{Big} to f_{small}: Subspace

$$
\begin{aligned}
f_{\mathsf{small}} &:= f_{\mathsf{Big}} \circ \boldsymbol{R} \\
f_{\mathsf{small}}(\boldsymbol{x}_{\mathsf{small}}) &:= f_{\mathsf{Big}}(\boldsymbol{R}\boldsymbol{x}_{\mathsf{Big}}) \\
\boldsymbol{x}_{\mathsf{small}} &:= \boldsymbol{R}\boldsymbol{x}_{\mathsf{Big}}
\end{aligned}
$$

- $\boldsymbol{R}$ is called *restriction* in PDEs
  - $\boldsymbol{R}$ is a short-wide matrix that maps from $\mathbb{R}^N$ to $\mathbb{R}^n$
  - if $n < N$ short-wide: not one-to-one but many-to-one
- $\boldsymbol{R}$ is a class of matrices
  - If $\boldsymbol{R}$ is the ith row of $\boldsymbol{I} \implies$ coordinate descent
  - If $\boldsymbol{R}$ is random $\implies$ random subspace method
  - If $\boldsymbol{R}$ comes from PDE $\implies$ multigrid method

How to go from f_{Big} to f_{small}: Subspace

$$\begin{aligned} f_{\text{small}} &:= f_{\text{Big}} \circ \boldsymbol{R} \\ f_{\text{small}}(\boldsymbol{x}_{\text{small}}) &:= f_{\text{Big}}(\boldsymbol{R}\boldsymbol{x}_{\text{Big}}) \\ \boldsymbol{x}_{\text{small}} &:= \boldsymbol{R}\boldsymbol{x}_{\text{Big}} \end{aligned}$$

- $\boldsymbol{R}$ is called *restriction* in PDEs
  - $\boldsymbol{R}$ is a short-wide matrix that maps from $\mathbb{R}^N$ to $\mathbb{R}^n$
  - if $n < N$ short-wide: not one-to-one but many-to-one
- $\boldsymbol{R}$ is a class of matrices
  - If $\boldsymbol{R}$ is the ith row of $\boldsymbol{I} \implies$ coordinate descent
  - If $\boldsymbol{R}$ is random $\implies$ random subspace method
  - If $\boldsymbol{R}$ comes from PDE $\implies$ multigrid method
- No-free-lunch
  - How to choose / design $\boldsymbol{R} \implies$ by experience, open problem
  - I am "cheating" by using a "known" $\boldsymbol{R}$

FAS multigrid (red are new) for $\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\boldsymbol{x}) + g(\boldsymbol{x})$

1. $\boldsymbol{x}_{\mathsf{Big}}^{k+1/2} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\mathsf{Big}}^{k} - \alpha\nabla f(\boldsymbol{x}_{\mathsf{Big}}^{k})\Big)$ ProxGD in Big world

FAS multigrid (red are new) for $\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\boldsymbol{x}) + g(\boldsymbol{x})$

1. $\boldsymbol{x}^{k+1/2}_{\text{Big}} = \text{prox}_{\alpha g}\Big(\boldsymbol{x}^{k}_{\text{Big}} - \alpha \nabla f(\boldsymbol{x}^{k}_{\text{Big}})\Big)$ ProxGD in Big world

2. $\boldsymbol{x}^{k+1/2}_{\text{small}} = \boldsymbol{R}\ \boldsymbol{x}^{k+1/2}_{\text{Big}}$ Restriction: Big-to-small

   $f_{\text{small}} = f_{\text{Big}} \circ \boldsymbol{R}, \quad g_{\text{small}} = g_{\text{Big}} \circ \boldsymbol{R}$

   adaptive $\boldsymbol{R}$ is new

FAS multigrid (red are new) for $\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + g(\boldsymbol{x})$

1. $\boldsymbol{x}_{\text{Big}}^{k+1/2} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\text{Big}}^{k} - \alpha\nabla f(\boldsymbol{x}_{\text{Big}}^{k})\Big)$ ProxGD in Big world

2. $\boldsymbol{x}_{\text{small}}^{k+1/2} = \boldsymbol{R}\,\boldsymbol{x}_{\text{Big}}^{k+1/2}$ Restriction: Big-to-small

   $f_{\text{small}} = f_{\text{Big}} \circ \boldsymbol{R}$, $g_{\text{small}} = g_{\text{Big}} \circ \boldsymbol{R}$

   adaptive $\boldsymbol{R}$ is new

3. $\boldsymbol{\tau} \in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$ WTF

   inclusion & Minkowski sum

FAS multigrid (red are new) for $\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\boldsymbol{x}) + g(\boldsymbol{x})$

1. $\boldsymbol{x}_{\text{Big}}^{k+1/2} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\text{Big}}^{k} - \alpha\nabla f(\boldsymbol{x}_{\text{Big}}^{k})\Big)$ ProxGD in Big world

2. $\boldsymbol{x}_{\text{small}}^{k+1/2} = \boldsymbol{R}\ \boldsymbol{x}_{\text{Big}}^{k+1/2}$ Restriction: Big-to-small

   $f_{\text{small}} = f_{\text{Big}} \circ \boldsymbol{R}, \quad g_{\text{small}} = g_{\text{Big}} \circ \boldsymbol{R}$

   adaptive $\boldsymbol{R}$ is new

3. $\boldsymbol{\tau} \in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$ WTF

   inclusion & Minkowski sum

4. $\boldsymbol{x}_{\text{small}}^{k+1} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\text{small}}^{k+1/2} - \alpha\big(\nabla f(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{\tau}\big)\Big)$ ProxGD in small world

FAS multigrid (red are new) for $\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\boldsymbol{x}) + g(\boldsymbol{x})$

1. $\boldsymbol{x}_{\text{Big}}^{k+1/2} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\text{Big}}^{k} - \alpha\nabla f(\boldsymbol{x}_{\text{Big}}^{k})\Big)$ ProxGD in Big world

2. $\boldsymbol{x}_{\text{small}}^{k+1/2} = \boldsymbol{R}\ \boldsymbol{x}_{\text{Big}}^{k+1/2}$ Restriction: Big-to-small

   $f_{\text{small}} = f_{\text{Big}} \circ \boldsymbol{R}, \quad g_{\text{small}} = g_{\text{Big}} \circ \boldsymbol{R}$

   adaptive $\boldsymbol{R}$ is new

3. $\boldsymbol{\tau} \in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$ WTF

   inclusion & Minkowski sum

4. $\boldsymbol{x}_{\text{small}}^{k+1} = \operatorname{prox}_{\alpha g}\Big(\boldsymbol{x}_{\text{small}}^{k+1/2} - \alpha\big(\nabla f(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{\tau}\big)\Big)$ ProxGD in small world

5. $\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + \alpha\boldsymbol{P}\ (\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$ use small to correct Big

   no prox

   existence of α is new

Where are the fun things

- Subdifferential can be $\varnothing$, crazy things will happen
  We fixed this: empty-set will not appear
- How to choose $\boldsymbol{\tau}$?
  We fixed this: the algo always works regardless of choice of $\boldsymbol{\tau}$
- $\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$ works?
  We fixed this: it always is a descent direction for $\boldsymbol{x}_{\text{Big}}$
- α exists?
  We fixed this: $\alpha > 0$ exists

So ... what is new?

$$\boldsymbol{\tau} \quad \in \quad \partial\Big(f_{\mathsf{small}}(\boldsymbol{x}_{\mathsf{small}}^{k+1/2}) + g_{\mathsf{small}}(\boldsymbol{x}_{\mathsf{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}}^{k+1/2}) + g_{\mathsf{Big}}(\boldsymbol{x}_{\mathsf{Big}}^{k+1/2})\Big) \qquad \text{Minkowski sum}$$

So ... what is new?

$$
\begin{aligned}
\boldsymbol{\tau} \quad &\in \quad \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Minkowski sum} \\
&= \quad \partial f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\partial f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) \oplus \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Moreau-Rockafellar thm. (new}^{1}\text{)}
\end{aligned}
$$

So ... what is new?

$$
\begin{aligned}
\boldsymbol{\tau} &\in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Minkowski sum} \\
&= \partial f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\partial f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) \oplus \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Moreau-Rockafellar thm. (new}^{1}\text{)} \\
&= \nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{diff.able part becomes singleton}
\end{aligned}
$$

So ... what is new?

$$\begin{aligned}
\boldsymbol{\tau} &\in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Minkowski sum} \\
&= \partial f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\partial f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) \oplus \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{Moreau-Rockafellar thm. (new}^1\text{)} \\
&= \nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) && \text{diff.able part becomes singleton} \\
&= \underbrace{\Big\{\nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{R}\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big\}}_{\text{singleton, simple part, exists in literature}} \; \underbrace{\oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus (-\boldsymbol{R})\partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})}_{\text{set Minkowski sum, hard part, new thing}}
\end{aligned}$$

So ... what is new?

$$\boldsymbol{\tau} \in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big) \ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) \quad \text{Minkowski sum}$$

$$= \partial f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\partial f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) \oplus \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) \quad \text{Moreau-Rockafellar thm. } (\text{new}^1)$$

$$= \nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \ominus \boldsymbol{R}\Big(\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + \partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big) \quad \text{diff.able part becomes singleton}$$

$$= \underbrace{\Big\{\nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{R}\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big\}}_{\text{singleton, simple part, exists in literature}} \underbrace{\oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) \oplus (-\boldsymbol{R})\partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})}_{\text{set Minkowski sum, hard part, new thing}}$$

$$= \underbrace{\Big\{\nabla f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{R}\nabla f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big\}}_{\text{singleton, simple part, exists in literature}} \oplus \partial g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + (-\boldsymbol{R})\partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) \quad \text{adaptive restriction } (\text{new}^2)$$

1. We made this true
2. To make my life easier: $\boldsymbol{R}$ maps all set-valued entries of $\partial g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})$ to the singleton $\{0\}$
   Open problem: keep dealing with $\oplus(-\boldsymbol{R})$ instead of $+(-\boldsymbol{R})$
3. the whole scheme is different from existing literature

What's the heck is $\boldsymbol{\tau}$?

- $\boldsymbol{\tau}$ links Big-world and small-world

What's the heck is $\boldsymbol{\tau}$?

- $\boldsymbol{\tau}$ links Big-world and small-world

- **Theorem** IF $\boldsymbol{x}_{\textsf{Big}}$ solves $\underset{\boldsymbol{x}_{\textsf{Big}}\in\mathbb{R}^n}{\operatorname{argmin}}\ f_{\textsf{Big}}(\boldsymbol{x}_{\textsf{Big}}) + g_{\textsf{Big}}(\boldsymbol{x}_{\textsf{Big}})$

  THEN $\boldsymbol{R}\boldsymbol{x}_{\textsf{Big}}$ solves $\underset{\boldsymbol{x}_{\textsf{small}}\in\mathbb{R}^n}{\operatorname{argmin}}\ f_{\textsf{small}}(\boldsymbol{x}_{\textsf{small}}) + g_{\textsf{small}}(\boldsymbol{x}_{\textsf{small}}) - \langle \boldsymbol{\tau}, \boldsymbol{x}_{\textsf{small}} \rangle$

What's the heck is $\boldsymbol{\tau}$?

- $\boldsymbol{\tau}$ links Big-world and small-world

- **Theorem** IF x_{Big} solves $\underset{\boldsymbol{x}_{\text{Big}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}})$

  THEN $\boldsymbol{R}\boldsymbol{x}_{\text{Big}}$ solves $\underset{\boldsymbol{x}_{\text{small}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{small}}(\boldsymbol{x}_{\text{small}}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}) - \langle \boldsymbol{\tau}, \boldsymbol{x}_{\text{small}} \rangle$

  (In algo) ProxGrad converges in Big-world $\iff$ ProxGrad converges in small-world

What's the heck is $\boldsymbol{\tau}$?

- $\boldsymbol{\tau}$ links Big-world and small-world

- **Theorem** IF $\boldsymbol{x}_{\text{Big}}$ solves $\underset{\boldsymbol{x}_{\text{Big}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}})$

  THEN $\boldsymbol{R}\boldsymbol{x}_{\text{Big}}$ solves $\underset{\boldsymbol{x}_{\text{small}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{small}}(\boldsymbol{x}_{\text{small}}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}) - \langle \boldsymbol{\tau}, \boldsymbol{x}_{\text{small}} \rangle$

  (In algo) ProxGrad converges in Big-world $\iff$ ProxGrad converges in small-world

- I am NOT saying: $\boldsymbol{x}_{\text{Big}}$ solves Big-problem $\iff$ $\boldsymbol{x}_{\text{small}}$ solves small-problem

- I am saying: $\boldsymbol{x}_{\text{Big}}$ solves Big-problem $\iff$ $\boldsymbol{x}_{\text{small}}$ solves small-perturbed-problem

- $\boldsymbol{\tau}$ is Big-perturb-small so that $\{$sol of small-problem$\}$ flavours $\{$sol of Big-problem$\}$

What's the heck is $\boldsymbol{\tau}$?

- $\boldsymbol{\tau}$ links Big-world and small-world

- **Theorem** IF $\boldsymbol{x}_{\text{Big}}$ solves $\underset{\boldsymbol{x}_{\text{Big}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}})$

  THEN $\boldsymbol{R}\boldsymbol{x}_{\text{Big}}$ solves $\underset{\boldsymbol{x}_{\text{small}} \in \mathbb{R}^n}{\operatorname{argmin}} f_{\text{small}}(\boldsymbol{x}_{\text{small}}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}) - \langle \boldsymbol{\tau}, \boldsymbol{x}_{\text{small}} \rangle$

  (In algo) ProxGrad converges in Big-world $\iff$ ProxGrad converges in small-world

- I am NOT saying: $\boldsymbol{x}_{\text{Big}}$ solves Big-problem $\iff$ $\boldsymbol{x}_{\text{small}}$ solves small-problem

- I am saying: $\boldsymbol{x}_{\text{Big}}$ solves Big-problem $\iff$ $\boldsymbol{x}_{\text{small}}$ solves small-perturbed-problem

- $\boldsymbol{\tau}$ is Big-perturb-small so that $\{$sol of small-problem$\}$ flavours $\{$sol of Big-problem$\}$

- How to prove: definition of $\boldsymbol{\tau}, \boldsymbol{R}$, convexity of obj functions, 1st-order subdiff. optimality

Coarse correction step: inequality

$$\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + \color{red}{\alpha}\color{black}\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$$

- **Theorem** $\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$ is a descent direction in Big-world

$$\left\langle \partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big),\ \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2}) \right\rangle \ < \ 0$$

the inequality is strict

Coarse correction step: inequality

$$\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + {\color{red}\alpha}\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$$

- **Theorem** $\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$ is a descent direction in Big-world

$$\left\langle \partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big),\ \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2}) \right\rangle \ < \ 0$$

  the inequality is strict

- The inequality holds for
  - **any** $\boldsymbol{\tau}$ you choose to get $\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau})$
  - **any** subgradient in $\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$

Coarse correction step: inequality

$$\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + {\color{red}\alpha}\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$$

- **Theorem** $\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$ is a descent direction in Big-world

$$\left\langle \partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big),\ \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2}) \right\rangle < 0$$

  the inequality is strict

- The inequality holds for
  - **any** $\boldsymbol{\tau}$ you choose to get $\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau})$
  - **any** subgradient in $\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$

- How to prove: definition & convexity

Coarse correction step: stepsize exists

$$\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + \alpha \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2}) \tag{A}$$

- **Theorem** $\alpha > 0$ exists such that $F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1}) < F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})$ $\qquad (F := f + g)$

- **3-sentence proof**
  1. The strict inequality in
  $$\left\langle \partial F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}),\ \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2}) \right\rangle < 0$$
  means $\partial F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})$ is strictly inside a half-space with normal $\boldsymbol{N} := \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2})$
  2. $\underbrace{\text{Subdifferential is a compact convex set}}_{\text{a fact}} \overset{1}{\Longrightarrow}$ strict separation $\Longrightarrow$ $\partial F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})$ must be a positive distance $(\alpha > 0)$ from that hyperplane defined by $\boldsymbol{N}$.
  3. Evaluate the support func. of $\partial F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})$, i.e., the directional derivative of F_{Big} at $\boldsymbol{x}_{\text{Big}}^{k+1/2}$ in the direction $\boldsymbol{N}$, we are done

- Sad news: we only have descent condition, not sufficient descent condition

- Deep thing in the compactness of subdifferential

Theoretical results

1. $\boldsymbol{x}_{\text{Big}}^{k+1/2} = \operatorname{prox}_{\alpha g}\left(\boldsymbol{x}_{\text{Big}}^{k} - \alpha \nabla f(\boldsymbol{x}_{\text{Big}}^{k})\right)$
2. $\boldsymbol{x}_{\text{small}}^{k+1/2} = \boldsymbol{R}\ \boldsymbol{x}_{\text{Big}}^{k+1/2}$
3. $\boldsymbol{\tau} \in \partial\Big(f_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2}) + g_{\text{small}}(\boldsymbol{x}_{\text{small}}^{k+1/2})\Big)$ $\ominus \boldsymbol{R}\partial\Big(f_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}) + g_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2})\Big)$
4. $\boldsymbol{x}_{\text{small}}^{k+1} = \operatorname{prox}_{\alpha g}\left(\boldsymbol{x}_{\text{small}}^{k+1/2} - \alpha(\nabla f(\boldsymbol{x}_{\text{small}}^{k+1/2}) - \boldsymbol{\tau})\right)$
5. $\boldsymbol{x}_{\text{Big}}^{k+1} = \boldsymbol{x}_{\text{Big}}^{k+1/2} + \alpha\boldsymbol{P}\ (\boldsymbol{x}_{\text{small}}^{k+1} - \boldsymbol{x}_{\text{small}}^{k+1/2})$

1. At convergence, $\boldsymbol{x}_{\ell}^{k}$ has a fixed-pt. property $\forall$ level ℓ
2. **Nonsmooth** angle condition
$$\left\langle \partial F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}),\ \boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2})\right\rangle < 0$$
3. Descent property: stepsize $\alpha > 0$ exists and $\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2})$ is a descent direction at $\boldsymbol{x}_{\text{Big}}^{k+1/2}$
$$F_{\text{Big}}\left(\boldsymbol{x}_{\text{Big}}^{k+1/2} + \alpha\boldsymbol{P}(\boldsymbol{x}_{\text{small}}^{k+1}(\boldsymbol{\tau}) - \boldsymbol{x}_{\text{small}}^{k+1/2})\right) < F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k+1/2}).$$
4. $\{F_0(\boldsymbol{x}_{\text{Big}}^{k})\}_{k\in\mathbb{N}}$ converges to $F_{\text{Big}}^{*} := \inf F_{\text{Big}}$, with
   - a sublinear rate
   $$F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k}) - F_{\text{Big}}^{*} \le \frac{c}{k}$$
   - a linear rate
   $$F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k}) - F_{\text{Big}}^{*} \le c\left(1 - \frac{\mu_{\text{Big}}}{L_{\text{Big}}}\right)^{k}.$$
   - with acceleration
   $$F_{\text{Big}}(\boldsymbol{x}_{\text{Big}}^{k}) - F_{\text{Big}}^{*} \le \frac{d}{ak^2 + bk + c}$$
5. $\{\boldsymbol{x}_{\text{Big}}^{k}\}_{k\in\mathbb{N}} \xrightarrow{k} \boldsymbol{x}_{\text{Big}}^{*}$

Why stop at 2-level?

Algorithm 4.1 L-level `MGProx` with V-cycle structure for an approximate solution of (1.1)

Initialize x_0^1 and the full version of $R_{\ell\to\ell+1}, P_{\ell+1\to\ell}$ for $\ell \in \{0,1,\ldots,L-1\}$
for $k = 1,2,\ldots$ **do**
  Set $\tau^{k+1}_{-1\to 0} = 0$
  for $\ell = 0,1,\ldots,L-1$ **do**
    $y_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell}g_\ell}\left(x_\ell^k - \dfrac{\nabla f_\ell(x_\ell^k) - \tau^{k+1}_{\ell-1\to\ell}}{L_\ell}\right)$ — pre-smoothing
    $x_{\ell+1}^k = R_{\ell\to\ell+1}(y_\ell^{k+1})\, y_\ell^{k+1}$ — restriction to next level
    $\tau^{k+1}_{\ell\to\ell+1} \in \underline{\partial F_{\ell+1}(x_{\ell+1}^k)} - R_{\ell\to\ell+1}(y_\ell^{k+1})\, \underline{\partial F_\ell(y_\ell^{k+1})}$ — create tau vector
  end for
  $w_L^{k+1} = \underset{\xi}{\operatorname{argmin}}\left\{ F_L^\tau(\xi) \coloneqq F_L(\xi) - \langle \tau^{k+1}_{L-1\to L}, \xi\rangle \right\}$ — solve the level-L coarse problem
  for $\ell = L-1, L-2, \ldots, 0$ **do**
    $z_\ell^{k+1} = y_\ell^{k+1} + \alpha P_{\ell+1\to\ell}(w_{\ell+1}^{k+1} - x_{\ell+1}^k)$ — coarse correction
    $w_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell}g_\ell}\left(z_\ell^{k+1} - \dfrac{\nabla f_\ell(z_\ell^{k+1}) - \tau^{k+1}_{\ell-1\to\ell}}{L_\ell}\right)$ — post-smoothing
  end for
  $x_0^{k+1} = w_0^{k+1}$ — update the fine variable
end for

ℓ	level
k	iteration counter
$\boldsymbol{x}_\ell$	variable at level ℓ before ProxGrad
$\boldsymbol{y}_\ell$	variable at level ℓ after ProxGrad
f_ℓ	smooth part at level ℓ
g_ℓ	nonsmooth part at level ℓ
L_ℓ	Lipschitz constant of ∇f_ℓ at level ℓ
L	number of levels

- Reduction in problem size $n_0 \to \dfrac{1}{4}n_0 \to \dfrac{1}{16}n_0 \to \dfrac{1}{64}n_0 \to \dfrac{1}{256}n_0 \to \dfrac{1}{1024}n_0$
- Per-iteration cost by geometric series $a + ar + ar^2 + \cdots \to \dfrac{a}{1-r}$. For $r = \dfrac{1}{4}$, V-cycle is $2.66n_0$ for all single proximal gradient update.

Experiment

- Tons of papers on multigrid methods for applications

Experiment

- Tons of papers on multigrid methods for applications
- Elastic Obstacle Problem:

$$\min_u \iint_\Omega \sqrt{1 + \| \nabla u \|_{L^2}^2} dxdy + \lambda \iint_\Omega \| (\phi - u)_+ \|_{L^1} dxdy \quad \text{s.t.} \quad u = 0 \text{ on } \partial\Omega,$$

$$\underset{\boldsymbol{u} \in \mathbb{R}^{N^2}}{\operatorname{argmin}} \; h^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{1 + \left(\boldsymbol{D}_{(i,j),:}\boldsymbol{u}\right)^2 + \left(\boldsymbol{E}_{(i,j),:}\boldsymbol{u}\right)^2} + h^2 \lambda \|(\boldsymbol{\phi} - \boldsymbol{u})_+\|_1$$

$$\underset{\boldsymbol{u} \in \mathbb{R}^{N^2}}{\operatorname{argmin}} \; f_0(\boldsymbol{u}) + g_0(\boldsymbol{u}) \coloneqq \sum_{i=1}^{N} \sum_{j=1}^{N} \psi\Big(\boldsymbol{F}_{(i,j),:}\boldsymbol{u}\Big) + \lambda \|(\boldsymbol{\phi} - \boldsymbol{u})_+\|_1, \tag{EOP}$$

$$\psi : \mathbb{R}^2 \to \mathbb{R} : (s,t) \mapsto \sqrt{1 + s^2 + t^2}, \qquad \boldsymbol{F}_{(i,j),:} \coloneqq \begin{bmatrix} \boldsymbol{D}_{(i,j),:} \\ \boldsymbol{E}_{(i,j),:} \end{bmatrix}.$$

Experiment

- Tons of papers on multigrid methods for applications
- Elastic Obstacle Problem:

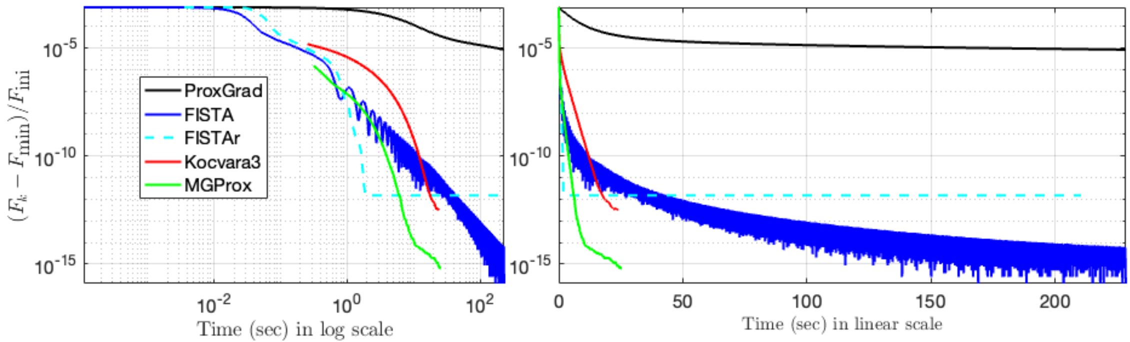
$$\min_u \iint_\Omega \sqrt{1 + \| \nabla u \|_{L^2}^2} dxdy + \lambda \iint_\Omega \| (\phi - u)_+ \|_{L^1} dxdy \quad \text{s.t.} \quad u = 0 \text{ on } \partial\Omega,$$

$$\underset{\boldsymbol{u}\in\mathbb{R}^{N^2}}{\operatorname{argmin}}\; h^2 \sum_{i=1}^N \sum_{j=1}^N \sqrt{1 + \big(\boldsymbol{D}_{(i,j),:}\boldsymbol{u}\big)^2 + \big(\boldsymbol{E}_{(i,j),:}\boldsymbol{u}\big)^2} + h^2\lambda\|(\boldsymbol{\phi} - \boldsymbol{u})_+\|_1$$

$$\underset{\boldsymbol{u}\in\mathbb{R}^{N^2}}{\operatorname{argmin}}\; f_0(\boldsymbol{u}) + g_0(\boldsymbol{u}) := \sum_{i=1}^N \sum_{j=1}^N \psi\Big(\boldsymbol{F}_{(i,j),:}\boldsymbol{u}\Big) + \lambda\|(\boldsymbol{\phi} - \boldsymbol{u})_+\|_1, \tag{EOP}$$

$$\psi : \mathbb{R}^2 \to \mathbb{R} : (s,t) \mapsto \sqrt{1 + s^2 + t^2}, \qquad \boldsymbol{F}_{(i,j),:} := \begin{bmatrix} \boldsymbol{D}_{(i,j),:} \\ \boldsymbol{E}_{(i,j),:} \end{bmatrix}.$$

- **Theorem** EOP problem is μ-strongly convex and smooth part is L-smooth
  - global tight L is unknown (to us)
  - $\mu > 0$ is unknown
  - $\mu \to 0$



(μ, L) unknown $\implies$ cannot do parameter-based restart

For $(2^8-1)^2 = 65025$ number of variable

Method	iterations k	time (sec.)	$\left(F_0(x_0^k) - F_0^{min}\right)/F_0(x_0^{\text{ini}})$
`ProxGrad`	$> 10^5$	335.9	8.33×10^{-8}
`FISTA`	$> 10^5$	332.62	6.64×10^{-8}
`FISTA-r`	$> 10^5$	364.89	6.64×10^{-8}
`Kocvara3` $N_s = 100$	$> 10^3$	986.53	8.07×10^{-8}
`MGProx` $N_s = 100$	50	48.37	1.32×10^{-10}

Last page - summary

Andersen Ang, Hans De Sterck, Steve Vavasis,
"MGProx: A nonsmooth multigrid proximal gradient method with adaptive restriction for strongly convex optimization",
SIAM Journal of Optimization, to appear, 2024
arXiv 2302.04077

- discussed some funny things
- many open problems / extensions / improvement opportunities
- Ads: I actually have PhD positions

End of document