MGProx: A nonsmooth **M**ulti**G**rid **Prox**imal gradient method, and +

Andersen Ang

ECS, Uni. Southampton, UK

andersen.ang@soton.ac.uk

Homepage angms.science

Version: June 16, 2023
First draft: Dec 2, 2022

Mathematical Optimization for Machine Learning
Humboldt-Universitat zu Berlin
June 14-16 2023

arXiv 2302.04077 joint work with

Hans De Sterck

Steve Vavasis

Standard setup in convex optimization

$$(\mathcal{P}) \quad : \quad \underset{x}{\operatorname{argmin}} \ \Big\{ F_0(x) := f_0(x) + g_0(x) \Big\}.$$

- $f_0 : \mathbb{R}^n \to \mathbb{R}$ convex, L-smooth[1] $\qquad f \in \mathcal{C}_L^{1,1}$
- $g_0 : \mathbb{R}^n \to \bar{\mathbb{R}}$ convex, possibly nonsmooth[2] $\qquad g$ cvx
  - $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ extended real

- To make (my) life easier:
  - Everything in finite dimensional Euclidean space $\qquad \mathbb{R}^n, \langle \cdot, \cdot \rangle, \| \cdot \|$
  - f_0 is strongly convex $\implies \mathcal{P}$ has an unique global sol $\qquad \operatorname{argmin} F_0$ is a singleton
  - g_0 is "proximable" $\implies \operatorname{prox}$ operator $\qquad \operatorname{prox}$ has closed-form / efficiently computable
  - F_0 has "multigrid-able" structure $\implies$ restriction, prolongation are given $\qquad R, P$ known
  - Assume all other necessary rigour things[3]

Topic today: solve $\mathcal{P}$ by proximal gradient method $\oplus$ multigrid.

---

[1] f_0 differentiable & ∇f_0 is L-Lipschitz

[2] not everywhere differentiable

[3] f_0 lower bounded, g_0 proper, lower-semicontinuous, lower level-bounded, prox-bounded with finite threshold, $\operatorname{prox}_{g_0}$ nonempty compact, f_0, g_0 both subdifferentiable

1 page review on solving $(\mathcal{P}) : \min \left\{ F_0(x) := f_0(x) + g_0(x) \right\}$

Proximal gradient iteration

$$
\begin{aligned}
x^+ &:= \text{prox}_{\alpha g_0}\Big(x - \alpha \nabla f_0(x)\Big) \\
&= \underset{\xi}{\text{argmin}} \;\; \alpha g_0(\xi) + \frac{1}{2}\Big\|\xi - \Big(x - \alpha \nabla f_0(x)\Big)\Big\|_2^2.
\end{aligned}
$$

- $\alpha \in (0, \frac{2}{L}]$ gradient stepsize. We fix $\alpha \equiv \frac{1}{L}$.
- prox operator of αg_0 at ζ:

$$\text{prox}_{\alpha g_0}(\zeta) := \underset{\xi}{\text{argmin}} \;\; \alpha g_0(\xi) + \frac{1}{2}\|\xi - \zeta\|_2^2.$$

  Usefulness: $\text{prox}_{\alpha g_0}$ fixes *nonsmoothness*
  - model regularization g_0
  - model constraint (indicator function) g_0

  Many $\text{prox}_{\alpha g_0}$ has closed-form sol.
- Literature history
  - Moreau envelope — Moreau 1962
  - Proximal point method — Rockafellar 1976
  - Forward-Backward splitting — Pasty 1979
  - Earliest proximal gradient — Fukushima & Mine 1981
  - Proximal FB splitting — Combettes & Wajs 2005
  - Now everywhere in Opt. & ML

Multigrid: coarse correction iteration

$$x^+ \;:=\; x + \alpha P(\hat{x}^+ - \hat{x}).$$

- Use coarse to improve fine
  - $\hat{x} \in \mathbb{R}^{n_1}$ restricted version of $x \in \mathbb{R}^{n_0}$
  - $\hat{x}^+$: obtained by solving an **auxiliary coarse optimization problem**, a "smaller" $\mathcal{P}$ (talk later)
  - P: prolongation
- History
  - For $g_0 \equiv 0$ (smooth convex optimization)
  - Linear system from the discretization of PDEs
  - Later generalized to system of nonlinear eqs
  - $\exists$ nonsmooth multigird in literature, but all different from this talk (see paper for detail)
- Usefulness: fast, convergence independent of problem size
- Literature history
  - Earliest(?) work on Poisson problem — Fedorenko 1962
  - Multi-level adaptive technique — Brandt 1973
  - Multigrid Methods — Hackbusch 1985
  - Now everywhere in scientific computing

This work

Proximal gradient

- ☺ Wide applications (due to g_0)
- ☹ Slow

Multigrid

- ☺ Fastest known method (at least for PDEs)
- ☹ Narrow applications: only for PDEs

Million dollar question: can we have both ☺?

`MGProx`: for some F_0, yes. 2022

`MGPD`: for more F_0, yes 2023

see arXiv 2302.04077 Section 1.4.2 for literature review

- Brandt & Cryer, Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems 1983
- Hackbusch & Mittelmann, On multi-grid methods for variational inequalities 1983
- Mandel, A multilevel iterative method for symmetric, positive definite linear complementarity problems 1984
- Vogel & Oman, Iterative methods for total variation denoising 1996
- Chan, Chan & Wana, Multigrid for differential-convolution problems arising from image processing 1998
- Nash, A multigrid approach to discretized optimization problems 2000
- Graser, Sack and Sander, Truncated nonsmooth Newton multigrid methods for convex minimization problems 2009
- Parpas, A multilevel proximal gradient algorithm for a class of composite optimization problems 2017
- Graser & Sander, Truncated nonsmooth Newton multigrid methods for block-separable minimization problems 2019

A remark on the popular MGOPT by Nash

Remark 1.1 (MGOPT has no theoretical convergence guarantee). The proof of [27, Theorem 1] on the convergence of MGOPT requires additional assumptions. In short the proof states the following: on solving (1.3) with an iterative algorithm $x^{k+1} := \sigma(x^k)$ where the update map $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be converging from any starting point x^1, now suppose $\rho : \mathbb{R}^n \to \mathbb{R}^n$ is some other operator with the descending property that $f_0(\rho(x)) \leq f_0(x)$. Then [27, Theorem 1] claimed that an algorithm consisting of interlacing σ with ρ repeatedly is also convergent. This is generally not true without further assumptions. E.g., consider a function $f(x_1, x_2)$ that is equal to $\frac{1}{1+x_2^2}$ on the set $U := \{(x_1, x_2) : |x_1| \geq 1\}$ and on the complementary set $\mathbb{R}^2 \setminus U$ that $f(x_1, x_2)$ has a unique minimizer at $(0, 0)$. Then $\sigma : (x_1, x_2) \mapsto \frac{9}{10}(x_1, x_2)$ and $\rho|_U : (x_1, x_2) \mapsto (\frac{10}{9}x_1, 2x_2)$ satisfies the hypothesis but diverges from any stationary point in $\{(x_1, x_2) : |x_1| \geq \frac{10}{9}\}$.

A first look at 2-level MGProx algorithm for $(\mathcal{P}) : \min_x \left\{F_0(x) \coloneqq f_0(x) + g_0(x)\right\}$

Algorithm 2.1 2-level MGProx for an approximate solu

Initialize x_0^1, R and P
for $k = 1, 2, \ldots$ **do**
(i) $y_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\left(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\right)$
(ii) $y_1^{k+1} = R(y_0^{k+1})y_0^{k+1}$
(iii) $\tau_{0\to1}^{k+1} \in \underline{\partial F_1(y_1^{k+1})} - R(y_0^{k+1})\,\underline{\partial F_0(y_0^{k+1})}$
(iv) $x_1^{k+1} = \underset{\xi}{\text{argmin}}\left\{F_1^\tau(\xi) \coloneqq F_1(\xi) - \langle \tau_{0\to1}^{k+1}, \xi\rangle\right\}$
(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$
(vi) $x_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\left(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\right)$
end for

- Variable sequence $\left\{x_0^k, y_0^k, z_0^k\right\}_{k\in\mathbb{N}}$
  - superscript k: iteration number
  - subscript 0: level
  - x: main sequence
  - y, z intermediate variables
- When converge: $x_0 = y_0 = z_0$ (fixed-point)

i prox-grad update at level-0
  - $\frac{1}{L_0}$ stepsize, L_0 is the Lipschitz const. of ∇f_0
  - this step is called "pre-smoothing" in multigrid
  - we use x to get y

ii **Adaptive restriction** of the updated y_0^{k+1}
  - R: (adaptive) restriction operator adapted to y_0^{k+1}

iii τ carries the level-0 info to level-1
  - ∂F_1: cvx subdifferential of F_1 at level 1
  - ∂F_0: cvx subdifferential of F_1 at level 0
  - τ can be any element of the set

iv Solve the coarse problem
  - a "smaller" $\mathcal{P}$ with a linear perturbation τ

v Coarse correction step
  - P: prolongate level-1 variable to level-0
  - we use x, y to get z

vi prox-grad update at level-0
  - $\frac{1}{L_0}$ stepsize, L_0 is the Lipschitz const. of ∇f_0
  - this step is called "post-smoothing" in multigrid
  - we use z to get x

Subdifferential, Minkowski sum and adaptive restriction

$$\text{(ii)} \qquad y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$$

$$\text{(iii)} \qquad \tau_{0\to 1}^{k+1} \in \underline{\tau_{0\to 1}^{k+1}} := \underline{\partial F_1(y_1^{k+1})} \oplus (-R)\underline{\partial F_0(y_0^{k+1})}$$

- **(Fenchel) Convex subdifferential** of a function $\phi : \mathbb{R}^n \to \mathbb{R}$ at a point x_0 is the set $\Big\{ \boldsymbol{q} \in \mathbb{R}^n \ : \ \phi(x) \geq \phi(x_0) + \langle \boldsymbol{q}, x - x_0 \rangle \Big\}$.
- Underline means set, no underline means singleton.
- Subdifferentials $\partial F_1(y_1^{k+1})$ & $\partial F_0(y_0^{k+1})$ are sets $\longrightarrow$ $\underline{\tau_{0\to 1}^{k+1}} := \underline{\partial F_1(y_1^{k+1})} \oplus (-R)\underline{\partial F_0(y_0^{k+1})}$ is a Minkowski sum.
- To make life easier, use R to turn $R\underline{\partial F_0(y_0^{k+1})}$ into a singleton vector.
  - R reduces $R\underline{\partial F_0(y_0^{k+1})}$ from a set-valued vector to a singleton vector. All sets map to the singleton $\{0\}$.
  - No more complicated Minkowski sum, now we have

$$\underline{\partial F_1(y_1^{k+1})} \oplus (-R)\underline{\partial F_0(y_0^{k+1})} \ = \ \underline{\partial F_1(y_1^{k+1})} - R\underline{\partial F_0(y_0^{k+1})}.$$

  - Not just "make life easier", the adaptive R plays critical role in proving convergence.
  - **Open problem**: non-adaptive R, general multi-member Minkowski sum of subdifferentials
- **Example for separable** g such as $\|\boldsymbol{x}\|_1$, $\max\{\boldsymbol{x}, \boldsymbol{c}\}$, etc.
  - **Definition** Let $\mathcal{I} = \Big\{ i \in [n] \ : \ [\partial F_0(y_0^{k+1})]_i \text{ is a set} \Big\}$.
  - **Adaptive restriction** R is defined as the (full) restriction matrix R_{full} with column $i \in \mathcal{I}$ set to zero.

Restriction and coarse level object

$$\text{(ii)} \qquad y_1^{k+1} = R(y_0^{k+1})y_0^{k+1}$$

$$\text{(iii)} \qquad \tau_{0\to 1}^{k+1} \in \underline{\partial F_1(y_1^{k+1})} - R\underline{\partial F_0(y_0^{k+1})}$$

$$\text{(iv)} \qquad x_1^{k+1} \in \underset{\xi}{\operatorname{argmin}} \left\{ F_1^\tau(\xi) \coloneqq F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi \rangle \right\}$$

- Level-0 variable $x_0 = Px_1$
- Level-1 variable $x_1 = Rx_0$
- Level-1 function $F_1(x_1) \coloneqq F_0(Px_1)$
- $F_1^\tau \coloneqq F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi \rangle$
- R, P preserve convexity

Example: 1-dimensional full weighting

$$R = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & & & \\ & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

maps vectors in $\mathbb{R}^{n_0}$ to $\mathbb{R}^{n_1}$ with $n_1 = \lceil \frac{n_0 - 1}{2} \rceil$.
$\implies 50\%$ reduction in problem size

$$P = 2R^\top$$

For 2-dimensional case, reduce size to $\dfrac{1}{4}$

Theoretical results

Algorithm 2.1 2-level MGProx for an approximate solu

Initialize x_0^1, R and P
for $k = 1, 2, \dots$ **do**
(i) $y_0^{k+1} = \text{prox}_{\frac{1}{L_0}g_0}\Big(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\Big)$
(ii) $y_1^{k+1} = R(y_0^{k+1})y_0^{k+1}$
(iii) $\tau_{0\to 1}^{k+1} \in \underline{\partial F_1(y_1^{k+1})} - R(y_0^{k+1})\,\underline{\partial F_0(y_0^{k+1})}$
(iv) $x_1^{k+1} = \underset{\xi}{\text{argmin}}\Big\{F_1^\tau(\xi) \coloneqq F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle\Big\}$
(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$
(vi) $x_0^{k+1} = \text{prox}_{\frac{1}{L_0}g_0}\Big(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\Big)$
end for

1. At convergence, x_ℓ^k has a fixed-pt. property $\forall \ell$
2. **Nonsmooth** angle condition $\Big\langle\, P(x_1^{k+1} - y_1^{k+1})\,,\, \partial F_0(y_0^{k+1})\,\Big\rangle \;<\; 0.$
3. Descent property: stepsize $\alpha > 0$ exists and $P(x_1^{k+1} - y_1^{k+1})$ is a descent direction at y_0^{k+1}

   i.e., $F_0\big(y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})\big) \;<\; F_0\big(y_0^{k+1}\big).$
4. $\big\{F_0(x_0^k)\big\}_{k\in\mathbb{N}}$ converges to $F_0^* \coloneqq \inf F_0$, with
   - a sublinear rate
     $$F_0(x_0^k) - F_0^* \leq \frac{\max\Big\{8\delta^2 L_0, F_0(x_0^1) - F_0^*\Big\}}{k}$$
     - L_0: Lipschitz constant of ∇f_0
     - δ: diameter of sublevel set $\{\boldsymbol{\xi} \in \mathbb{R}^{n_0} \mid F_0(\boldsymbol{\xi}) \leq F_0(\boldsymbol{x}_0^1)\}$
   - a linear rate
     $$F_0(x_0^k) - F^* \leq \Big(1 - \frac{\mu}{L_0}\Big)^k \big(F_0(x_1^k) - F^*\big).$$

   Both holds so
   $$F_0(x_0^k) - F^* \leq \min\Big\{\,\frac{\text{const.}}{k}\,,\,\Big(1 - \frac{\mu}{L_0}\Big)^k\,\Big\}.$$
5. $\{\boldsymbol{x}_0^k\}_{k\in\mathbb{N}} \xrightharpoonup{k} \boldsymbol{x}_0^*$

How we prove them

1. At convergence, x_ℓ^k has a fixed-pt. property $\forall \ell$
2. **Nonsmooth** angle condition
$$\Big\langle\, P(x_1^{k+1} - y_1^{k+1})\,,\, \partial F_0(y_0^{k+1})\,\Big\rangle \;<\; 0.$$
3. Descent property: stepsize $\alpha > 0$ exists and $P(x_1^{k+1} - y_1^{k+1})$ is a descent direction at y_0^{k+1}
$$\text{i.e., } F_0\big(y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})\big) \;<\; F_0\big(y_0^{k+1}\big).$$
4. $\big\{F_0(x_0^k)\big\}_{k\in\mathbb{N}}$ converges to $F_0^* := \inf F_0$, with
   - a sublinear rate
   $$F_0(x_0^k) - F_0^* \le \frac{\max\Big\{8\delta^2 L_0, F_0(x_0^1) - F_0^*\Big\}}{k}$$
   - a linear rate
   $$F_0(x_0^k) - F^* \le \Big(1 - \frac{\mu}{L_0}\Big)^k \big(F_0(x_1^k) - F^*\big).$$

   Both holds so
   $$F_0(x_0^k) - F^* \le \min\Big\{\, \frac{\text{const.}}{k}\,,\, \Big(1 - \frac{\mu}{L_0}\Big)^k\,\Big\}.$$
5. $\{\boldsymbol{x}_0^k\}_{k\in\mathbb{N}} \xrightarrow{k} \boldsymbol{x}_0^*$

1.
   - Fixed-pt. property of proximal gradient step
   - Adaptive R reduces set to singleton
   - Subgradient 1st-order optimality
2.
   - Adaptive R reduces set to singleton
   - Definition of τ and x_1^{k+1}
   - Convexity of F_1
   - Restriction preserves convexity
3.
   - Result 2 (angle condition)
   - Subdifferential ∂F is a compact convex set
   - Strict hyperplane separation
   - Support of ∂F = directional derivative of F
4.
   - Result 3 (descent property) & 4 lemmas
     - a sufficient "descent" inequality
     - a quadratic overestimator of F_0
     - diameter of sublevel set of F_0
     - an inequality of scalar sequence

     & a bunch of convex analysis techniques
   - Result 3 (descent property) & the proximal Polyak-Łojasiewics inequality

   Both convergences results are **global** (regardless of starting pt.)
5. Result 4 and F_0 is strictly convex by assumption

Fixed-point property

Algorithm 2.1 2-level MGProx for an approximate solu

Initialize x_0^1, R and P
for $k = 1, 2, \dots$ **do**
(i) $y_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\Big)$
(ii) $y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$
(iii) $\tau_{0\to 1}^{k+1} \in \partial F_1(y_1^{k+1}) - R(y_0^{k+1})\,\partial F_0(y_0^{k+1})$
(iv) $x_1^{k+1} = \underset{\xi}{\text{argmin}}\Big\{F_1^\tau(\xi) := F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle\Big\}$
(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$
(vi) $x_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\Big)$
end for

Theorem 2.5 (Fixed-point). *In Algorithm 2.1, if x_0^k solves (1.1), then we have the fixed-point properties $x_0^{k+1} = y_0^{k+1} = x_0^k$ and $x_1^{k+1} = y_1^{k+1}$.*

Proof. The fixed-point property of the proximal gradient operator [32, page 150] gives

(2.6) $$y_0^{k+1} \overset{\text{fixed-point}}{=} x_0^k \overset{\text{assumption}}{=} \text{argmin}\; F_0(x).$$

As a result, the coarse variable satisfies

(2.7) $$y_1^{k+1} := R y_0^{k+1} \overset{(2.6)}{=} R x_0^k,$$

The subgradient 1st-order optimality to $y_0^{k+1} \overset{(2.6)}{\in} \text{argmin}\; F_0(x)$ gives $0 \in \partial F_0(y_0^{k+1})$. Multiplying by $-R$ (which reduces the set $\partial F_0(x_0^k)$ to a singleton) gives

(2.8) $$0 = -R\partial F_0(x_0^k).$$

Then adding $\partial F_1(y_1^{k+1})$ on both sides of (2.8) gives

(2.9a) $$\partial F_1(y_1^{k+1}) = \partial F_1(y_1^{k+1}) - R(x_0^k)\partial F_0(x_0^k)$$
(2.9b) $$\overset{(2.4a)}{\ni} \tau_{0\to 1}^{k+1}$$

In (2.8), $-R\partial F_0(x_0^k)$ is the zero vector, so the equality in (2.9a) holds since we are adding zero to a (non-empty) set. The inclusion (2.9b) follows from (2.4a) as $\partial F_1(y_1^{k+1}) - R(x_0^k)\partial F_0(x_0^k)$ is the set $\tau_{0\to 1}^{k+1}$.

Now rearranging (2.9b) gives $0 \in \partial F_1(y_1^{k+1}) - \tau_{0\to 1}^{k+1}$, which is exactly the subgradient 1st-order optimality condition for the coarse problem $\underset{\xi}{\text{argmin}}\, F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle$. By strong convexity of $F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle$, the point y_1^{k+1} is the unique minimizer of the coarse problem, so $x_1^{k+1} = y_1^{k+1}$ by step (iv) of the algorithm and $x_0^{k+1} = y_0^{k+1} \overset{(2.6)}{=} x_0^k$ by steps (v) and (vi). □

Nonsmooth angle condition

Algorithm 2.1 2-level `MGProx` for an approximate solu

Initialize x_0^1, R and P

for $k = 1, 2, \ldots$ **do**

(i) $y_0^{k+1} = \operatorname{prox}_{\frac{1}{L_0} g_0}\Big(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\Big)$

(ii) $y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$

(iii) $\tau_{0\to 1}^{k+1} \in \partial F_1(y_1^{k+1}) - R(y_0^{k+1})\, \partial F_0(y_0^{k+1})$

(iv) $x_1^{k+1} = \operatorname*{argmin}_{\xi} \Big\{ F_1^\tau(\xi) := F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle \Big\}$

(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$

(vi) $x_0^{k+1} = \operatorname{prox}_{\frac{1}{L_0} g_0}\Big(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\Big)$

end for

Theorem 2.6 (Angle condition of coarse correction). *For $P(x_1^{k+1} - y_1^{k+1}) \neq 0$, the following directional derivative is strictly negative*

$$\Big\langle \partial F_0(y_0^{k+1}), P(x_1^{k+1} - y_1^{k+1}) \Big\rangle < 0. \tag{2.10}$$

Before we prove the theorem we emphasize that (2.10) applies for any subgradient in the set $\partial F_0(y_0^{k+1})$. Furthermore,

$$(2.10) \iff \Big\langle P^\top \partial F_0(y_0^{k+1}), x_1^{k+1} - y_1^{k+1} \Big\rangle < 0 \overset{P^\top = cR,\, c>0}{\iff} c\Big\langle R\partial F_0(y_0^{k+1}), x_1^{k+1} - y_1^{k+1} \Big\rangle < 0.$$

As c, R, P are all element-wise nonnegative, showing (2.10) is equivalent to showing

$$\Big\langle R\partial F_0(y_0^{k+1}), x_1^{k+1} - y_1^{k+1} \Big\rangle < 0, \tag{2.11}$$

where $R\partial F_0(y_0^{k+1})$ is a singleton vector for all subgradients in $\partial F_0(y_0^{k+1})$ due to the adaptive R.

Proof. By definition $\tau_{0\to 1}^{k+1} \overset{(2.4a)}{\in} \partial F_1(y_1^{k+1}) - R\partial F_0(y_0^{k+1})$ hence

$$R\partial F_0(y_0^{k+1}) \in \partial F_1(y_1^{k+1}) - \tau_{0\to 1}^{k+1} \overset{(2.5)}{=} \partial F_1^\tau(y_1^{k+1}), \tag{2.12}$$

showing that $R\partial F_0(y_0^{k+1})$ is a subgradient of F_1^τ at y_1^{k+1}. For any subgradient in the subdifferential $\partial F_1^\tau(y_1^{k+1})$, we have the following which implies (2.11):

$$\Big\langle \partial F_1^\tau(y_1^{k+1}), x_1^{k+1} - y_1^{k+1} \Big\rangle < F_1^\tau(x_1^{k+1}) - F_1^\tau(y_1^{k+1}) < 0,$$

where the first strict inequality is due to F_1^τ being a strongly convex function (which implies strict convexity) ; the second inequality is by $x_1^{k+1} := \operatorname*{argmin}_{\xi} F_1^\tau(\xi)$ and the assumption that $x_1^{k+1} \neq y_1^{k+1}$. ∎

Remark 2.7. Theorem 2.6 holds for convex but not strongly convex f_0 by replacing $<$ with $\leq$.

Descent property

Algorithm 2.1 2-level `MGProx` for an approximate solu

Initialize x_0^1, R and P

for $k = 1, 2, \ldots$ **do**

(i) $y_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\Big)$

(ii) $y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$

(iii) $\tau_{0\to 1}^{k+1} \in \partial F_1(y_1^{k+1}) - R(y_0^{k+1})\, \partial F_0(y_0^{k+1})$

(iv) $x_1^{k+1} = \underset{\xi}{\text{argmin}} \left\{ F_1^\tau(\xi) := F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle \right\}$

(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$

(vi) $x_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\Big)$

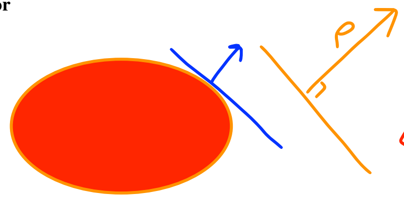
end for

Lemma 2.8 (Existence of stepsize). *There exists* $\alpha_k > 0$ *such that* (2.13) *is satisfied for* $P(x_1^{k+1} - y_1^{k+1}) \neq 0$.

To prove the lemma, we make use the second definition of subdifferential we discussed in subsection 2.2: $\partial F_0(y_0^{k+1})$ is a compact convex set whose support function is the directional derivative of F_0 at y_0^{k+1}. Note that $F_0 : \mathbb{R}^{n_0} \to \overline{\mathbb{R}}$ will never reach $+\infty$ at z_0^{k+1} since z_0^{k+1} is obtained by the proximal gradient step, so we can make use of the result on directional derivative in [19, Def. 1.1.4, p.165] associated with subdifferential.

Proof. We prove the lemma in 3 steps.

1. (Halfspace) The strict inequality in Theorem 2.6 means that $\partial F_0(y_0^{k+1})$ is strictly inside a halfspace with normal vector $p = P(x_1^{k+1} - y_1^{k+1})$.
2. (Strict separation) Being a compact convex set, $\partial F_0(y_0^{k+1}0)$ lying strictly on one side of the hyperplane must be a positive distance (say $\alpha_k > 0$) from that hyperplane.
3. (Support and directional derivative) Evaluating the support function of $\partial F_0(y_0^{k+1})$, i.e., the directional derivative of F_0 at y_0^{k+1} in the direction p, we have (2.13). ∎



$$\langle \partial F_0(y_0^{k+1}), P(x_1^{k+1} - y_1^{k+1}) \rangle < 0$$

Sublinear rate convergence

Algorithm 2.1 2-level MGProx for an approximate solu

Initialize x_0^1, R and P

for $k = 1, 2, \dots$ **do**

(i) $y_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\Big)$

(ii) $y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$

(iii) $\tau_{0\to 1}^{k+1} \in \partial F_1(y_1^{k+1}) - R(y_0^{k+1})\,\partial F_0(y_0^{k+1})$

(iv) $x_1^{k+1} = \underset{\xi}{\text{argmin}} \Big\{ F_1^\tau(\xi) := F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi \rangle \Big\}$

(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$

(vi) $x_0^{k+1} = \text{prox}_{\frac{1}{L_0} g_0}\Big(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\Big)$

end for

- Existing proof framework of prox-grad method cannot be used.
- MGProx is interlacing two update operations
- Prox-grad iteration guarantee descent of function value

$$f(\xi^+) \leq f\big(\text{ProxGradUpdate}(\xi)\big) \quad (*)$$

- descent of function value does not imply variable getting closer to sol.

$$(*) \;\not\Longrightarrow\; \|\xi^+ - \xi^*\| \leq \|\xi - \xi^*\|$$

Lemma 2.11 (Sufficient descent of MGProx iteration). *For all iterations k, we have*

$$F(x^{k+1}) - F^* \leq \frac{L}{2}\Big(\|x^k - x^*\|_2^2 - \|y^{k+1} - x^*\|_2^2\Big). \tag{2.15}$$

Lemma 2.13 (A quadratic overestimator). *For all x, we have*

$$F(x) - F(x^{k+1}) \geq L\langle x^k - y^{k+1}, x - x^k \rangle + \frac{L}{2}\|y^{k+1} - x^k\|_2^2. \tag{2.19}$$

Lemma 2.14 (Diameter of sublevel set). *At initial guess $x^1 \in \mathbb{R}^n$, define*

$$\mathcal{L}_{\leq F(x^1)} := \Big\{ x \in \mathbb{R}^n \mid F(x) \leq F(x^1) \Big\}, \qquad \textit{(sublevel set of } x^1\textit{)}$$

$$\delta = \textit{diam}\; \mathcal{L}_{\leq F(x^1)} := \sup\Big\{ \|x - y\|_2 \mid F(x) \leq F(x^1), F(y) \leq F(y^1) \Big\}. \qquad \textit{(diameter of } \mathcal{L}_{\leq F(x^1)}\textit{)}$$

Then for $x^* :=$ argmin $F(x)$, we have $\|x^k - x^*\|_2 \leq \delta$ and $\|y^k - x^*\|_2 \leq \delta$ for all k.

Proof. We have $F(x^*) \leq F(x^1)$ by definition. By the descent property of the coarse correction and proximal gradient updates, we have $F(x^k) \leq F(x^1)$ and $F(y^k) \leq F(x^1)$ for all k. These results mean that x^k, y^{k+1} and x^* are inside $\mathcal{L}_{\leq F(x^1)}$, therefore both $\|x^k - x^*\|_2$ and $\|y^{k+1} - x^*\|_2$ are bounded above by δ. Lastly, F is strongly convex so $\mathcal{L}_{\leq F(x^1)}$ is bounded and $\delta < +\infty$. □

Lemma 2.15 (Monotone sequence). *For a nonnegative sequence $\{\omega_k\}_{k\in\mathbb{N}} \to \omega^*$ that is monotonically decreasing with $\omega_1 - \omega^* \leq 4\mu$ and $\omega_k - \omega_{k+1} \geq \frac{(\omega_{k+1}-\omega^*)^2}{\mu}$, it holds that $\omega_k - \omega^* \leq \frac{4\mu}{k}$ for all k.*

Proof. By induction. See proof in [22, Lemma 4]. □

Lemma 2.11 + Lemma 2.13 + Lemma 2.14 + Lemma 2.15 = sublinear rate

$$F_0(x_0^k) - F_0^* \leq \frac{\text{const.}}{k}$$

Linear rate convergence via proximal Polyak-Łojasiewics inequality

Algorithm 2.1 2-level `MGProx` for an approximate solu

Initialize x_0^1, R and P

for $k = 1, 2, \dots$ **do**

(i) $y_0^{k+1} = \mathrm{prox}_{\frac{1}{L_0} g_0}\left(x_0^k - \frac{1}{L_0}\nabla f(x_0^k)\right)$

(ii) $y_1^{k+1} = R(y_0^{k+1}) y_0^{k+1}$

(iii) $\tau_{0\to 1}^{k+1} \in \partial F_1(y_1^{k+1}) - R(y_0^{k+1})\,\partial F_0(y_0^{k+1})$

(iv) $x_1^{k+1} = \underset{\xi}{\mathrm{argmin}}\left\{F_1^\tau(\xi) := F_1(\xi) - \langle \tau_{0\to 1}^{k+1}, \xi\rangle\right\}$

(v) $z_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$

(vi) $x_0^{k+1} = \mathrm{prox}_{\frac{1}{L_0} g_0}\left(z_0^{k+1} - \frac{1}{L_0}\nabla f(z_0^{k+1})\right)$

end for

2.4.6. Linear convergence rate by Proximal PŁ inequality. All the functions and variables here are at level 0 so we omit the subscripts. Now we show that $\{F(x^k)\}_{k\in\mathbb{N}}$ converges to F^* with a linear rate using the *Proximal Polyak-Łojasiewics inequality* [21, Section 4]. The function F in Problem (1.1) is called ProxPŁ if there exists $\mu > 0$ such that

$$\text{(ProxPŁ)} \qquad \frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(F(x) - F^*) \qquad \forall x,$$

where μ is called the ProxPŁ constant and

$$\text{(2.25)} \qquad \mathcal{D}_g(x,\alpha) := -2\alpha \min_z \left\{\frac{\alpha}{2}\|z-x\|_2^2 + \langle z-x, \nabla f(x)\rangle + g(z) - g(x)\right\}.$$

Intuitively, $\mathcal{D}_g$ is defined based on the proximal gradient operator:

$$\mathrm{prox}_{\frac{1}{L}g}\left(x - \frac{\nabla f(x)}{L}\right) \overset{(2.21)}{=} \underset{z}{\mathrm{argmin}}\ \frac{L}{2}\|z-x\|_2^2 + \langle z-x, \nabla f(x)\rangle + g(z) - g(x).$$

It has been shown in [21] that if f in (1.1) is μ-strongly convex, then F is μ-ProxPŁ. Now we

Theorem 2.16. *Let x_0^1 be the initial guess of the algorithm, $F_0^* = F_0(x_0^*)$ and $x_0^* = \mathrm{argmin}\ F_0(x)$. The sequence $\{x_0^k\}_{k\in\mathbb{N}}$ generated by* `MGProx` *(Algorithm 2.1) for solving Problem (1.1) satisfies $F_0(x_0^{k+1}) - F_0^* \leq \left(1 - \frac{\mu_0}{L_0}\right)^k\left(F_0(x_0^1) - F_0^*\right)$.*

Parameters in the algorithm

- Gradient stepsize in the proximal gradient iteration $y_0^{k+1} = \text{prox}_{\alpha g}\Big(x_0^k - \alpha \nabla f(x_0^k)\Big)$

  just use constant stepsize $\alpha = \dfrac{1}{L_0}$

- The selection of τ in $\underline{\tau_{0\to 1}^{k+1}} \in \underline{\partial F_1(y_1^{k+1})} - R\underline{\partial F_0(y_0^{k+1})}$

  any possible τ in the set $\underline{\tau}$ is ok

- Coarse correction stepsize in $y_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$

  just use any naive line search on α for $F_0\Big(y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})\Big) \; < \; F_0\Big(y_0^{k+1}\Big)$

  - $<$ becomes $=$ when $x_1^{k+1} = y_1^{k+1}$, .i.e., we reached fixed-pt. (convergence).
  - We deal with nonsmooth problem, cannot use classical stuffs like Armijo rule, Wolfe condition, Goldstein line search: they assume function F_0 is differentiable
  - We do not need sufficient descent condition for `MGProx` because the sufficient descent condition from proximal gradient iteration is sufficient
  - Design line search with nonsmooth sufficient descent condition is possible, but out of scope.
    In fact, line search for nonsmooth descent is very deep, linked to the Kurdyka-Łojasiewicz inequality.

Algorithm 3.1 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**

$$y_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell} g_\ell}\left(x_\ell^k - \frac{\nabla f_\ell(x_\ell^k) - \tau^{k+1}_{\ell-1\to\ell}}{L_\ell} \right) \quad \text{pre-smoothing}$$

$$x_{\ell+1}^k = R_{\ell\to\ell+1}(y_\ell^{k+1})\, y_\ell^{k+1} \quad \text{restriction to next level}$$

$$\tau^{k+1}_{\ell\to\ell+1} \in \underline{\partial F_{\ell+1}(x_{\ell+1}^k)} - R_{\ell\to\ell+1}(y_\ell^{k+1})\, \underline{\partial F_\ell(y_\ell^{k+1})} \quad \text{create tau vector}$$

end for

$$w_L^{k+1} = \operatorname*{argmin}_{\xi} \left\{ F_L^\tau(\xi) \coloneqq F_L(\xi) - \langle \tau^{k+1}_{L-1\to L}, \xi \rangle \right\} \quad \text{solve the level-}L\text{ coarse problem}$$

for $\ell = L-1, L-2, \ldots, 0$ **do**

$$z_\ell^{k+1} = y_\ell^{k+1} + \alpha P_{\ell+1\to\ell}(w_{\ell+1}^{k+1} - x_{\ell+1}^k) \quad \text{coarse correction}$$

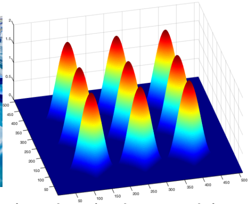
$$w_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell} g_\ell}\left(z_\ell^{k+1} - \frac{\nabla f_\ell(z_\ell^{k+1}) - \tau^{k+1}_{\ell-1\to\ell}}{L_\ell} \right) \quad \text{post-smoothing}$$

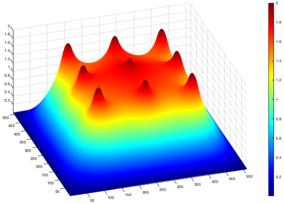
end for

$$x_0^{k+1} = w_0^{k+1} \quad \text{update the fine variable}$$

end for

Elastic Obstacle Problem $\min_{u \geq \phi} \int_\Omega \sqrt{1 + \|\nabla u\|_{L^2}^2} dxdy \approx \min_{u \geq \phi} \int_\Omega \frac{1}{2} \|\nabla u\|_{L^2}^2 dxdy$





- Given obstacle ϕ, find a membrane $u \geq \phi$ with the min. elastic potential energy.

$\min_u \int_\Omega \frac{1}{2} \|\nabla u\|_{L^2}^2 dxdy$	minimum variation	$\Omega \subset \mathbb{R}^2$	domain
s.t. $u \geq \phi$, in Ω	obstacle constraint	$\phi(x,y) : \mathbb{R}^2 \to \mathbb{R}$	obstacle
$u = 0$, on $\partial\Omega$	boundary condition	$u(x,y) : \mathbb{R}^2 \to \mathbb{R}$	membrane
		$\nabla u : \mathbb{R}^2 \to \mathbb{R}^2$	gradient field of u

- N-by-N grid discretization:

$$\min_{u \in \mathbb{R}^{N^2}} \underbrace{\frac{1}{2}\langle Q_0 u, u\rangle}_{f_0} + \underbrace{i_{\geq \phi}(u)}_{g_0}, \quad Q := \frac{1}{h^2}\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \approx \nabla^2, \quad i_{\geq \phi}(u) = \begin{cases} 0 & u \geq \phi \\ \infty & u < \phi \end{cases}$$

- Why this problem: $\because$ people know what R, P can be used.
- Can we use `MGProx` on other problem: yes if you give me the R, P that will work.

cost to pay

On $\min_x \left\{ F_0(x) := \frac{1}{2}\langle Q_0 x, x\rangle + i_{\geq\phi}(x) \right\}$

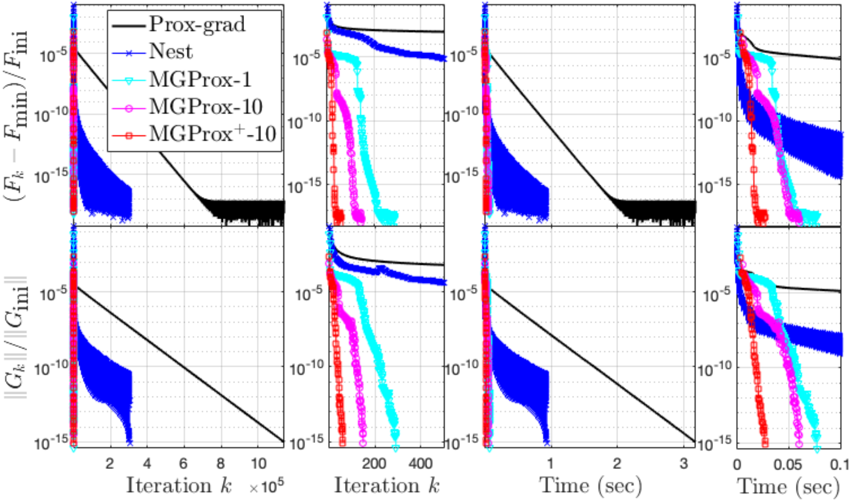


FIGURE 2. *Typical convergence plots of* `Prox`*,* `Nest`*,* `MGProx-1`*,* `MGProx-10` *and* `MGProx`$^+$`-10` *for 1-dimensional (Shifted aEOP). The number of variables in this experiment is* $2^9 - 1 = 511$*. All* `MGProx` *methods use 7 levels.*

Different Elastic Obstacle Problems

$$\min_x \left\{ F_0(x) := f_0(x) + g_0(x) \right\}.$$

- Previous slide: Constrained approximated EOP

$$f_0(x) = \frac{1}{2}\langle Q_0 x, x\rangle, \quad g_0(x) = i_{\geq\phi}(x)$$

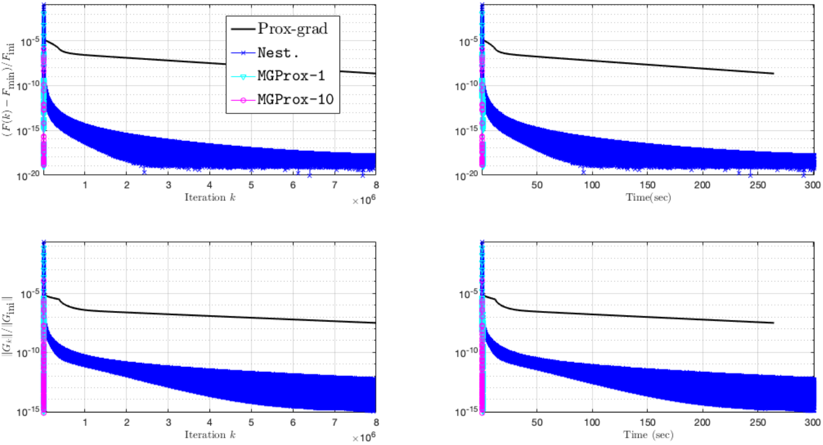
- Now: Unconstrained penalized approximated EOP

$$f_0(x) = \frac{1}{2}\langle Q_0 x, x\rangle, \quad g_0(x) = \mu\|(\phi - u)_+\|_1.$$

- Unconstrained penalized full EOP

$$f_0(x) = \sqrt{1 + \langle Q_0 x, x\rangle}, \quad g_0(x) = \mu\|(\phi - u)_+\|_1.$$

On $\min_x \left\{ F_0(x) := \frac{1}{2}\langle Q_0 x, x\rangle + \mu\|(\phi - u)_+\|_1 \right\}$

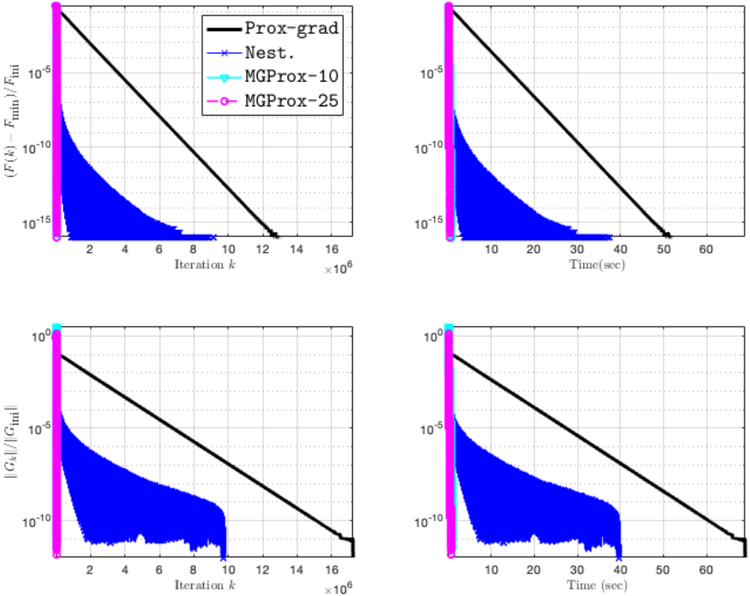


Run time

MGProx: $<$ 1sec reach 10^{-15}

Nesterov & Prox-grad:
not yet converge after 300sec

On $\min_x \left\{ F_0(x) := \sqrt{1 + \langle Q_0 x, x \rangle} + \mu \|(\phi - u)_+\|_1 \right\}$



Num iteration

MGProx: 10^2 reach 10^{-15}

Nesterov: 10^6

Prox-grad: 10^7

Run time

MGProx: < 1sec

Nesterov: 40sec

Prox-grad: 70sec

Why so fast?

- The coarse correction

$$x_0^{k+1} = y_0^{k+1} + \alpha P(x_1^{k+1} - y_1^{k+1})$$

- Reduction in problem size

$$n_0 \to \frac{1}{4}n_0 \to \frac{1}{16}n_0 \to \frac{1}{64}n_0 \to \frac{1}{256}n_0 \to \frac{1}{1024}n_0$$

- Per-iteration cost by geometric series $a, r \in (0,1)$

$$a + ar + ar^2 + \cdots \to \frac{a}{1-r}.$$

  For $n = \dfrac{1}{4}$ gives $1.33n_0$. V-cycle is then $2.66n_0$ for all single proximal gradient update.

- Can you add Nesterov's acceleration to MGProx?
  - No. In fact Nesterov's acceleration works very badly with MGProx.
    Why: due to Nesterov's ripples in the convergence.
    However, you can add Nesterov's acceleration in the pre/post-smoothing iteation.

Other things / future works

- Theory
  - Grid independence: convergence rate is independent of problem size
  - Classical Fourier analysis of multigrid
- Algorithms
  - `MGProx` that also corrects the active points
  - `MGProx` on proximal averages
  - Multigrid Proximal (quasi) Newton's method
  - **Nonsmooth multigrid trust-region method**
  - **Nonsmooth multigrid ADMM**
  - Nonsmooth multigrid manifold optimization
  - Block nonconvex but bi-convex problems (matrix factorizations)
- Applications
  - Image deblurring, dezooming, completion
  - Volumetric imaging (e.g. 3D medical imaging)
  - PDE-based image processing
  - Graphs

Last page - summary

- Multigrid proximal gradient method
- Adaptive restriction
- Theoretical characterizations
  - Fixed-pt
  - Angle and descent condition
  - Existence of line search stepsize
  - Global sublinear convergence rate
  - Global linear convergence rate
- Fast in experiments

Algorithm 3.1 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**

$$y_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell} g_\ell}\left(x_\ell^k - \frac{\nabla f_\ell(x_\ell^k) - \tau_{\ell-1\to\ell}^{k+1}}{L_\ell} \right) \quad \text{pre-smoothing}$$

$$x_{\ell+1}^k = R_{\ell\to\ell+1}(y_\ell^{k+1})\, y_\ell^{k+1} \quad \text{restriction to next level}$$

$$\tau_{\ell\to\ell+1}^{k+1} \in \underline{\partial F_{\ell+1}(x_{\ell+1}^k)} - R_{\ell\to\ell+1}(y_\ell^{k+1})\, \underline{\partial F_\ell(y_\ell^{k+1})} \quad \text{create tau vector}$$

  end for

$$w_L^{k+1} = \operatorname*{argmin}_\xi \left\{ F_L^\tau(\xi) := F_L(\xi) - \langle \tau_{L-1\to L}^{k+1}, \xi \rangle \right\} \quad \text{solve the level-}L\text{ coarse problem}$$

  for $\ell = L-1, L-2, \dots, 0$ **do**

$$z_\ell^{k+1} = y_\ell^{k+1} + \alpha P_{\ell+1\to\ell}(w_{\ell+1}^{k+1} - x_{\ell+1}^k) \quad \text{coarse correction}$$

$$w_\ell^{k+1} = \operatorname{prox}_{\frac{1}{L_\ell} g_\ell}\left(z_\ell^{k+1} - \frac{\nabla f_\ell(z_\ell^{k+1}) - \tau_{\ell-1\to\ell}^{k+1}}{L_\ell} \right) \quad \text{post-smoothing}$$

  end for

$$x_0^{k+1} = w_0^{k+1} \quad \text{update the fine variable}$$

end for

Paper arXiv2302.04077 now under review. Slide available angms.science

End of document

Primal-dual extension (New!)

- A non-diagonal evil $\boldsymbol{A}$ will make proximal gradient method does not work well on

$$\operatorname{argmin} f(\boldsymbol{x}) + g(\boldsymbol{A}\boldsymbol{x}).$$

- Convex-concave primal-dual problem

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \operatorname*{argmax}_{\boldsymbol{\lambda}\in\mathbb{R}^m} L(\boldsymbol{x}, \boldsymbol{\lambda})$$

  - Component-wise subgradient $\mathcal{D} := \begin{pmatrix} \partial_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \\ -\partial_{\boldsymbol{\lambda}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \end{pmatrix}$
  - Subdifferential 1st-order optimality condition

$$\mathbf{0} \in \begin{pmatrix} \partial_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \\ -\partial_{\boldsymbol{\lambda}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \end{pmatrix} + \boldsymbol{W} \begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k \end{pmatrix}$$

  - Chambolle-Pock Primal-dual hybrid gradient is $\boldsymbol{W} = \begin{pmatrix} \frac{1}{\eta}\boldsymbol{I} & \boldsymbol{A}^\top \\ \boldsymbol{A} & \frac{1}{\eta}\boldsymbol{I} \end{pmatrix}$
  - ADMM is $\boldsymbol{W} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \eta\boldsymbol{A}^\top\boldsymbol{A} & -\boldsymbol{A}^\top \\ \mathbf{0} & -\boldsymbol{A} & \frac{1}{\eta}\boldsymbol{I} \end{pmatrix}$

Algorithm 1: 2-level `MGPD`

Input: L

Output: $\boldsymbol{z}^k$ that approximately solve (1)

Initialize $\boldsymbol{z}^1, \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{P}$

for $k = 1, 2, \ldots$ **do**

Get $\boldsymbol{z}_0^{k+\frac{1}{3}}$ via solving the inclusion % pre-smoothing at level-0

$$\mathbf{0} \;\in\; \mathcal{D}_0(\boldsymbol{z}_0^{k+\frac{1}{3}}) + \boldsymbol{W}\big(\boldsymbol{z}_0^{k+\frac{1}{3}} - \boldsymbol{z}_0^k\big)$$

Block-wise coarsification % coarsification

$$\boldsymbol{z}_1^{k+\frac{1}{3}} \;=\; \mathcal{R}(\boldsymbol{z}_0^{k+\frac{1}{3}}) \;:=\; \begin{pmatrix} \boldsymbol{R}_1 & \\ & \boldsymbol{R}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_0^{k+\frac{1}{3}} \\ \boldsymbol{\lambda}_0^{k+\frac{1}{3}} \end{pmatrix}$$

Tau: % tau vecotr

$$\boldsymbol{\tau}_{0\to 1}^{k+1} \in \mathcal{D}_1(\boldsymbol{z}_1^{k+\frac{1}{3}}) - \mathcal{R}\mathcal{D}_0(\boldsymbol{z}_0^{k+\frac{1}{3}}) = \begin{pmatrix} \partial_{\boldsymbol{x}_1} L_1\big(\boldsymbol{x}_1^{k+\frac{1}{3}}, \boldsymbol{\lambda}_1^{k+\frac{1}{3}}\big) \\ \partial_{\boldsymbol{z}_1} L_1\big(\boldsymbol{x}_1^{k+\frac{1}{3}}, \boldsymbol{\lambda}_1^{k+\frac{1}{3}}\big) \end{pmatrix} - \begin{pmatrix} \boldsymbol{R}_1 & \\ & \boldsymbol{R}_2 \end{pmatrix} \begin{pmatrix} \partial_{\boldsymbol{x}_0} L_0\big(\boldsymbol{x}_0^{k+\frac{1}{3}}, \boldsymbol{\lambda}_0^{k+\frac{1}{3}}\big) \\ \partial_{\boldsymbol{z}_0} L_0\big(\boldsymbol{x}_0^{k+\frac{1}{3}}, \boldsymbol{\lambda}_0^{k+\frac{1}{3}}\big) \end{pmatrix}$$

Solve the coarse problem % solve the level-1 coarse problem

$$\boldsymbol{z}_1^{k+\frac{2}{3}} \in \operatorname*{argmin}_{\boldsymbol{x}_1} \operatorname*{argmax}_{\boldsymbol{\lambda}_1} L_1(\boldsymbol{x}_1, \boldsymbol{\lambda}_1) + \langle \boldsymbol{\tau}_{0\to 1}^{k+1}, \boldsymbol{z}_1 \rangle \;=\; L_1(\boldsymbol{x}_1, \boldsymbol{\lambda}_1) + \left\langle \begin{pmatrix} {}^1\boldsymbol{\tau}_{0\to 1}^{k+1} \\ {}^2\boldsymbol{\tau}_{0\to 1}^{k+1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{\lambda}_1 \end{pmatrix} \right\rangle$$

Coarse correction % Coarse correction

$$\boldsymbol{z}_0^{k+\frac{2}{3}} = \boldsymbol{z}_0^{k+\frac{1}{3}} + \begin{pmatrix} a & -\alpha \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_1 & \\ & \boldsymbol{P}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1^{k+\frac{2}{3}} - \boldsymbol{x}_1^{k+\frac{1}{3}} \\ \boldsymbol{\lambda}_1^{k+\frac{2}{3}} - \boldsymbol{\lambda}_1^{k+\frac{1}{3}} \end{pmatrix}$$

Get $\boldsymbol{z}_0^{k+1}$ via solving the inclusion % post-smoothing at level-0

$$\mathbf{0} \in \mathcal{D}_0(\boldsymbol{z}_0^{k+1}) + \boldsymbol{W}\big(\boldsymbol{z}_0^{k+1} - \boldsymbol{z}_0^{k+\frac{2}{3}}\big)$$

Now repeat the poof of MGProx on MGPD

“mind-blown.gif”

END OF PDF

(New New !)

Algorithm 1: FMGProx: Fast MGProx with Nesterov's acceleration

Input: The constants L of f

Output: x^k the approximately solve (1)

Initialization $z^0 = x^0, \gamma^0 > 0$

for $k = 1, 2, \ldots$ **do**

Compute $\alpha^k \in \left]0, 1\right[$ from $L(\alpha^k)^2 = (1-\alpha^k)\gamma^k$ `// extrapolation parameter`

$$\gamma^{k+1} = (1-\alpha^k)\gamma^k$$ `// extrapolation parameter`

$$y^k \quad = \alpha^k z^k + (1-\alpha^k)x^k$$ `// Nesterov's extrapolation`

$$x^{k+1} = \Big(\text{MGProx-V-cycle} \circ \text{prox}_{\frac{1}{L}g}\Big)\Big(y^k - \frac{1}{L}\nabla f(y^k)\Big)$$ `// prox-grad step with MGProx V-cycle`

$$g^k \quad = \frac{y^k - x^{k+1}}{L}$$ `// a ''gradient''`

$$z^{k+1} = z^k - \frac{\alpha^k}{\gamma^{k+1}} g^k$$ `// updating the auxiliary sequence`

Lemma 1. *Assuming*

$$F(x_k^*(y^k)) \ \le \ M_k\big(x_k^*(y^k); y^k\big) \tag{A0}$$

$$f \text{ is } L\text{-smooth and } \mu\text{-strongly convex,} \tag{A1}$$

$$\phi^0(x) \text{ is a convex function,} \tag{A2}$$

$$\big\{y^k\big\} \text{ is an arbitrary sequence,} \tag{A3}$$

$$\big\{\alpha^k\big\} \text{ is a sequence that } \alpha^k \in \,]\,0,1\,[, \tag{A4a}$$

$$\big\{\alpha^k\big\} \text{ is a sequence that } \sum_{k=0}^{\infty} \alpha^k = \infty, \tag{A4b}$$

$$\lambda^0 \ := \ 1 \tag{A5}$$

$$\lambda^{k+1} \ := \ (1-\alpha^k)\lambda^k \tag{A6}$$

$$\phi^{k+1}(x) \ := \ (1-\alpha^k)\phi^k(x) + \alpha^k\Big[F\big(x_k^*(y^k)\big) + \big\langle g^k, x - y^k\big\rangle + \frac{1}{2L}\|g^k\|_2^2\Big] \tag{A7}$$

Then the pair of sequences $\{\phi^k(x), \lambda^k\}$ *generated as in* (A6), (A7) *is an estimate sequence of* F.

Lemma 2. *Let* $\phi^0(x) \coloneqq F(x^0) + \frac{\gamma^0}{2}\|x - z^0\|_2^2$. *Then* ϕ^{k+1} *generated recursively as in* (A7) *in Lemma 1 has a closed-form expression*

$$\phi^{k+1}(x) \;=\; \overline{\phi}^{k+1} + \frac{\gamma^{k+1}}{2}\|x - z^{k+1}\|_2^2, \tag{8}$$

where

$$\gamma^{k+1} = (1 - \alpha^k)\gamma^k, \tag{9a}$$

$$z^{k+1} = z^k - \frac{\alpha^k}{\gamma^{k+1}} g^k, \tag{9b}$$

$$\overline{\phi}^{k+1} = (1 - \alpha^k)\overline{\phi}^k + \alpha^k F\big(x_k^*(y^k)\big) + \frac{\alpha^k}{2}\Big(\frac{1}{L} - \frac{\alpha^k}{\gamma^{k+1}}\Big)\|g^k\|_2^2 + \alpha^k\langle g^k, z^k - y^k\rangle. \tag{9c}$$

Lemma 3. *For minimization problem* (1), *assume* $x^* \in X^* \coloneqq \text{argmin } F(x)$ *exists and denote* $F^* \coloneqq F(x^*)$. *Suppose* $F(x^k) \le \overline{\phi}^k \coloneqq \min_x \phi_k(x)$ *holds for a sequence* $\{x^k\}_{k\in\mathbb{N}}$, *where* $\{\phi^k, \lambda^k\}_{k\in\mathbb{N}}$ *is an estimate sequence of* F, *and we define* $\phi^0 \coloneqq F(x^0) + \frac{\gamma^0}{2}\|x^0 - x^*\|_2^2$, *then we have for all* $k \in \mathbb{N}$ *that*

$$F(x^k) - F^* \;\le\; \lambda^k\Big[F(x^0) + \frac{\gamma^0}{2}\|x^0 - x^*\|_2^2 - F^*\Big].$$

Theorem 1. *Suppose* $F(x^k) \leq \overline{\phi}^k := \min_x \phi_k(x)$ *holds for a sequence* $\{x^k\}_{k\in\mathbb{N}}$*, where* $\{\phi^k, \lambda^k\}_{k\in\mathbb{N}}$ *is an estimate sequence of* F*. Define* $\phi^0 := F(x^0) + \frac{\gamma^0}{2}\|x^0 - x^*\|_2^2$*. Assuming all the conditions in Lemma 1, Lemma 2 and Lemma 3. Then we have*

$$0 \; < \; \lambda^k \; < \; \frac{4L}{(1-\alpha^k)\left(\gamma^0 k^2 + 4\sqrt{\gamma^0 L}k + 4L\right)}.$$

Corollary 1. *For the sequence* $\{x^k\}$ *produced by Algorithm FMGProx, we have*

$$F(x^k) - F^* \; \leq \; \frac{4L}{(1-\alpha^k)\left(\gamma^0 k^2 + 4\sqrt{\gamma^0 L}k + 4L\right)}\left[F(x^0) + \frac{\gamma^0}{2}\|x^0 - x^*\|_2^2 - F^*\right].$$

$$\leq \frac{\text{const.}}{k^2} \quad \text{(optimal)}$$