九州大学学術情報リポジトリ
Kyushu University Institutional Repository

ADAPTIVE PROXIMAL GRADIENT METHODS FOR CONVEX BILEVEL OPTIMIZATION

Themelis, Andreas
Kyushu University

Latafat, Puya
KU Leuven

Villa, Silvia
Università di Genova

Patrinos, Panagiotis
KU Leuven

https://hdl.handle.net/2324/6790346

出版情報：2023-05
バージョン：
権利関係：

ADAPTIVE PROXIMAL GRADIENT METHODS FOR CONVEX BILEVEL OPTIMIZATION

Andreas Themelis

andreas.themelis@ees.kyushu-u.ac.jp

Kyushu University

Puya Latafat

puya.latafat@kuleuven.be

KU Leuven

Silvia Villa

villa@dima.unige.it

Università di Genova

Panagiotis Patrinos

panos.patrinos@esat.kuleuven.be

KU Leuven

Control & Optimisation Pisa 2023

May 8-10, 2023

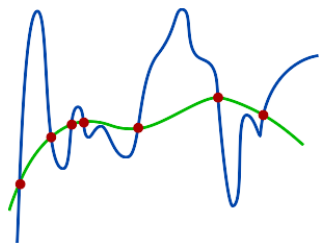
Outline

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Pick the "best" among solutions
  - "simplest"
  - minimum norm
  - (group-) sparsest
  - least noisy



- Origins in game theory *(Stackelberg competition)*

$$\begin{aligned} &\text{minimize} && \Phi^{(1)}(x, y) && \textit{leader} \\ &\text{subject to} && x \in \arg\min \Phi^{(2)}(\,\cdot\,, y) && \textit{follower} \end{aligned}$$

- (simple BP) includes many optimization setups (more on this later)

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Hard version of regularization

$$\text{minimize } \sigma \underbrace{\varphi^{(1)}}_{\textit{regularizer}} + \overbrace{\varphi^{(2)}}^{\textit{loss}}$$

as $\sigma \searrow 0$

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Hard version of regularization

$$\text{minimize } \sigma \underbrace{\varphi^{(1)}}_{\textit{regularizer}} + \overbrace{\varphi^{(2)}}^{\textit{loss}}$$

as $\sigma \searrow 0$

- **Example:** sparse solutions to linear system

$$\begin{aligned} &\text{minimize} && \sigma\|x\|_1 + \|Ax - b\|^2 \\ &\text{subject to} && x \in \mathbb{R}^n \end{aligned} \qquad \text{Lasso}$$

$$\begin{aligned} &\text{minimize} && \|x\|_1 \\ &\text{subject to} && Ax = b \quad (\text{requires } b \in \operatorname{range} A) \end{aligned} \qquad \text{(N)LP}$$

$$\begin{aligned} &\text{minimize} && \|x\|_1 \\ &\text{subject to} && x \in \arg\min \|A \cdot - b\|^2 \end{aligned} \qquad \text{Bilevel}$$

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Hard version of regularization

$$\text{minimize } \sigma \underbrace{\varphi^{(1)}}_{\textit{regularizer}} + \overbrace{\varphi^{(2)}}^{\textit{loss}}$$

  as $\sigma \searrow 0$

- Suggests penalty-type approach:
  1. choose $\sigma^+ < \sigma$
  2. find $x^+ \in \arg\min \sigma^+ \varphi^{(1)} + \varphi^{(2)}$
  3. $(\sigma, x) \leftarrow (\sigma^+, x^+)$

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Hard version of regularization

$$\text{minimize } \sigma \underbrace{\varphi^{(1)}}_{\textit{regularizer}} + \overbrace{\varphi^{(2)}}^{\textit{loss}}$$

  as $\sigma \searrow 0$

- Suggests penalty-type approach:
  1. choose $\sigma^+ < \sigma$
  2. find $x^+ \in \arg\min \sigma^+ \varphi^{(1)} + \varphi^{(2)}$ HARD!
  3. $(\sigma, x) \leftarrow (\sigma^+, x^+)$

Bilevel optimization

Goals

$$\begin{aligned} &\text{minimize} && \varphi^{(1)}(x) \\ &\text{subject to} && x \in \arg\min \varphi^{(2)} \end{aligned} \qquad \text{(simple BP)}$$

- Hard version of regularization

$$\text{minimize } \sigma \underbrace{\varphi^{(1)}}_{\textit{regularizer}} + \overbrace{\varphi^{(2)}}^{\textit{loss}}$$

  as $\sigma \searrow 0$

- Suggests penalty-type approach:
  1. choose $\sigma^+ \leqslant \sigma$ (not too small...)
  2. ~~find $x^+ \in \arg\min \sigma^+ \varphi^{(1)} + \varphi^{(2)}$~~ $x^+ = \operatorname{ProxGrad}_{\sigma^+ \varphi^{(1)} + \varphi^{(2)}}(x)$
  3. $(\sigma, x) \leftarrow (\sigma^+, x^+)$

✓ This talk: (simple BP) **without inner solvers**

- Throughout, everything convex

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005

[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005

[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014

[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓
iterative-3D[4]	$C^{1,1}$	✗	$C^{1,1}$		✓

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005

[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014

[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017

[4] G. Garrigos, L. Rosasco and S. Villa, *"Iterative regularization via dual diagonal descent"*, J. Math. Imag. Vision 60:189–215, 2018

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓
iterative-3D[4]	$C^{1,1}$	✗	$C^{1,1}$		✓
Solodov[5]	$C^{1,+}$	✗	$C^{1,+}$	δ_D	✓

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005
[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014
[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017
[4] G. Garrigos, L. Rosasco and S. Villa, *"Iterative regularization via dual diagonal descent"*, J. Math. Imag. Vision 60:189–215, 2018
[5] M. Solodov, *"An explicit descent method for bilevel convex optimization"*, J. of Conv. Anal. 14(2):227, 2007

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓
iterative-3D[4]	$C^{1,1}$	✗	$C^{1,1}$		✓
Solodov[5]	$C^{1,+}$	✗	$C^{1,+}$	δ_D	✓
this talk → adaBiM	$C^{1,+}$		$C^{1,+}$		✓

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005
[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014
[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017
[4] G. Garrigos, L. Rosasco and S. Villa, *"Iterative regularization via dual diagonal descent"*, J. Math. Imag. Vision 60:189–215, 2018
[5] M. Solodov, *"An explicit descent method for bilevel convex optimization"*, J. of Conv. Anal. 14(2):227, 2007

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓
iterative-3D[4]	$C^{1,1}$	✗	$C^{1,1}$		✓
Solodov[5]	$C^{1,+}$	✗	$C^{1,+}$	δ_D	✓
this talk → adaBiM	$C^{1,+}$		$C^{1,+}$		✓

- Solodov & adaBiM both *"adaptive"*: estimate *local* Lipschitz moduli

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005
[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014
[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017
[4] G. Garrigos, L. Rosasco and S. Villa, *"Iterative regularization via dual diagonal descent"*, J. Math. Imag. Vision 60:189–215, 2018
[5] M. Solodov, *"An explicit descent method for bilevel convex optimization"*, J. of Conv. Anal. 14(2):227, 2007

Bilevel optimization

Algorithmic literature

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Proximal gradient structure: $f^{(i)}$ smooth, $\sigma g^{(1)} + g^{(2)}$ "prox-friendly"

	upper level $\varphi^{(1)}$		lower level $\varphi^{(2)}$		explicit?
	$f^{(1)}$	$g^{(1)}$	$f^{(2)}$	$g^{(2)}$	
Cabot[1]	✗		✗		✗
MNG[2]	C^1, str. cvx	✗	$C^{1,1}$	δ_D	✗
BiGSAM[3]	$C^{1,1}$, str. cvx	✗	$C^{1,1}$, str. cvx		✓
iterative-3D[4]	$C^{1,1}$	✗	$C^{1,1}$		✓
Solodov[5]	$C^{1,+}$	✗	$C^{1,+}$	δ_D	✓
this talk → adaBiM	$C^{1,+}$		$C^{1,+}$		✓
this talk → staBiM	$C^{1,1}$		$C^{1,1}$		✓

- Solodov & adaBiM both *"adaptive"*: estimate *local* Lipschitz moduli
- adaBiM *"**fully** adaptive"*: insensitive to param. initialization (more on this later)

[1] A. Cabot, *"Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization"*, SIAM J. Opt. 15(2):555–572, 2005
[2] A. Beck and S. Sabach, *"A first order method for finding minimal norm-like solutions of convex optimization problems"*, Math. Prog. 147(1-2):25–46, 2014
[3] S. Sabach and S. Shtern, *"A First Order Method for Solving Convex Bilevel Optimization Problems"*, SIAM J. Opt. 27(2):640-660, 2017
[4] G. Garrigos, L. Rosasco and S. Villa, *"Iterative regularization via dual diagonal descent"*, J. Math. Imag. Vision 60:189–215, 2018
[5] M. Solodov, *"An explicit descent method for bilevel convex optimization"*, J. of Conv. Anal. 14(2):227, 2007

Bilevel optimization

Examples: primal-dual splitting

In the previous talk: primal-dual splitting methods

$$\text{minimize } f(x) + g(x) + h(Ax)$$

- ✓ $f \in C^{1+}$ (e.g. $f \in C^2$)
- ▶ g, h "prox-friendly"
- ✓ linesearch-free (traditionally, either $f \in C^{1,1}$ or LS needed)

C^{1+}: *locally* Lipschitz differentiable functions
$C^{1,1}$: *globally* " " "

Bilevel optimization

Examples: primal-dual splitting

In the previous talk: primal-dual splitting methods

$$\text{minimize } f(x) + g(x) + h(Ax)$$

- ✓ $f \in C^{1+}$ (e.g. $f \in C^2$)
- ▶ g, h "prox-friendly"
- ✓ linesearch-free (traditionally, either $f \in C^{1,1}$ or LS needed)

Bilevel reformulation

$$\begin{aligned} &\text{minimize } f(x) + g(x) + h(z) \\ &\text{subject to } (x, z) \in \operatorname*{arg\,min}_{u,v} \|Au - v\|^2 \end{aligned}$$

C^{1+}: *locally* Lipschitz differentiable functions
$C^{1,1}$: *globally* " " "

Bilevel optimization

Examples: primal-dual splitting

In the previous talk: primal-dual splitting methods

$$\text{minimize } f(x) + g(x) + h(Ax)$$

- ✓ $f \in C^{1+}$ (e.g. $f \in C^2$)
- ▶ g, h "prox-friendly"
- ✓ linesearch-free (traditionally, either $f \in C^{1,1}$ or LS needed)

Bilevel reformulation

$$\begin{aligned} &\text{minimize } \overbrace{f(x)}^{f^{(1)}} + \overbrace{g(x) + h(z)}^{g^{(1)}} \\ &\text{subject to } (x, z) \in \arg\min_{u,v} \underbrace{\|Au - v\|^2}_{f^{(2)}} \end{aligned}$$

C^{1+}: *locally* Lipschitz differentiable functions
$C^{1,1}$: *globally* " " "

Bilevel optimization

Examples: nonlinear programs

Nonsmooth nonlinear programs:

$$\begin{aligned} &\text{minimize } f(x) + g(x) \\ &\text{subject to } Ax = b,\ h(x) \leqslant 0 \end{aligned}$$

- ✓ $f, h \in C^{1+}$ (e.g. $f, h \in C^2$)
- ▶ g "prox-friendly"

Bilevel optimization

Examples: nonlinear programs

Nonsmooth nonlinear programs:

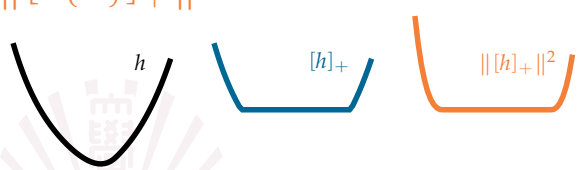
$$\begin{aligned} &\text{minimize } f(x) + g(x) \\ &\text{subject to } Ax = b,\ h(x) \leqslant 0 \end{aligned}$$

- ✓ $f, h \in C^{1+}$ (e.g. $f, h \in C^2$)
- ▶ g "prox-friendly"

Bilevel reformulation

$$\begin{aligned} &\text{minimize } f(x) + g(x) \\ &\text{subject to } x \in \arg\min_u \|Au - b\|^2 + \|[h(x)]_+\|^2 \end{aligned}$$

$\|[h(x)]_+\|^2$ convex and $C^{1,+}$



$[h(x)]_+ := \max\{h(x), 0\}$

Bilevel optimization

Examples: nonlinear programs

Nonsmooth nonlinear programs:

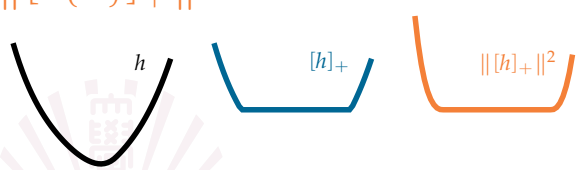
$$\begin{aligned} &\text{minimize } f(x) + g(x) \\ &\text{subject to } Ax = b,\ h(x) \leqslant 0 \end{aligned}$$

- ✓ $f, h \in C^{1+}$ (e.g. $f, h \in C^2$)
- ▶ g "prox-friendly"

Bilevel reformulation

$$\begin{aligned} &\text{minimize } \overbrace{f(x)}^{f^{(1)}} + \overbrace{g(x)}^{g^{(1)}} \\ &\text{subject to } x \in \arg\min_u \underbrace{\|Au - b\|^2 + \|[h(x)]_+\|^2}_{f^{(2)}} \end{aligned}$$

$\|[h(x)]_+\|^2$ convex and $C^{1,+}$



$[h(x)]_+ := \max\{h(x), 0\}$

Outline

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

2. $x^k = \Pi_D\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

2. $x^k = \Pi_D\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$
   $\alpha_k = \alpha_{\max}\eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$f_k(x^k) \leqslant f_k(x^{k-1}) + \nu\left\langle \nabla f_k(x^{k-1}), x^k - x^{k-1}\right\rangle$$

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

2. $x^k = \Pi_D\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$
   $\alpha_k = \alpha_{\max}\eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$f_k(x^k) \leqslant f_k(x^{k-1}) + \nu\left\langle \nabla f_k(x^{k-1}), x^k - x^{k-1}\right\rangle$$

   - $\nu, \eta \in (0,1)$ and $\alpha_{\max}$ fixed at initialization


An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

2. $x^k = \Pi_D\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$
   $\alpha_k = \alpha_{\max}\eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$f_k(x^k) \leqslant f_k(x^{k-1}) + \nu\left\langle \nabla f_k(x^{k-1}), x^k - x^{k-1}\right\rangle$$

   - $\nu, \eta \in (0,1)$ and $\alpha_{\max}$ fixed at initialization
   - ✗ cost evaluations ✓ no gradient evaluations

An adaptive proximal gradient solver

Precursors

Solodov's method ($f^{(1)}, f^{(2)} \in C^{1,+}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) \\ &\text{subject to } x \in \arg\min_{u \in D} f^{(2)}(u) \end{aligned}$$

Iteration k

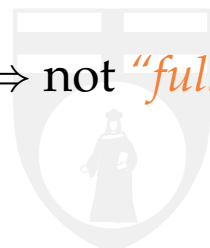
1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Idea:** projected gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} \quad \text{subject to } x \in D$$

2. $x^k = \Pi_D\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$
   $\alpha_k = \alpha_{\max}\eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$f_k(x^k) \leqslant f_k(x^{k-1}) + \nu\left\langle \nabla f_k(x^{k-1}), x^k - x^{k-1}\right\rangle$$

$\alpha_{\max}$ fixed $\Rightarrow$ not *"fully"* adaptive



An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Same idea:** *proximal* gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} + \underbrace{\sigma_k g^{(1)}(x) + g^{(2)}(x)}_{g_k(x)}$$

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Same idea:** *proximal* gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} + \underbrace{\sigma_k g^{(1)}(x) + g^{(2)}(x)}_{g_k(x)}$$

2. $x^k = \operatorname{prox}_{\alpha_k g_k}\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Same idea:** *proximal* gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} + \underbrace{\sigma_k g^{(1)}(x) + g^{(2)}(x)}_{g_k(x)}$$

2. $x^k = \operatorname{prox}_{\alpha_k g_k}\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$

   $\alpha_k = \overline{\alpha}_k \eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$\alpha_k \frac{\left\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1} \right\rangle}{\|x^k - x^{k-1}\|^2} \leqslant \nu$$

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

Iteration k

1. choose *inverse penalty* $\sigma_k \leqslant \sigma_{k-1}$
   - **Same idea:** *proximal* gradient step on penalized problem

$$\text{minimize } \underbrace{\sigma_k f^{(1)}(x) + f^{(2)}(x)}_{f_k(x)} + \underbrace{\sigma_k g^{(1)}(x) + g^{(2)}(x)}_{g_k(x)}$$

2. $x^k = \operatorname{prox}_{\alpha_k g_k}\left(x^{k-1} - \alpha_k \nabla f_k(x^{k-1})\right)$

   $\alpha_k = \overline{\alpha}_k \eta^{m_k}$, with $m_k \in \mathbb{N}$ the smallest such that

$$\alpha_k \frac{\left\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1} \right\rangle}{\|x^k - x^{k-1}\|^2} \leqslant \nu$$

   - $\nu, \eta \in (0,1)$ fixed at initialization, $\overline{\alpha}_k$ "fully" adapting to *local geometry*

✗ gradient evaluations ✓ no cost evaluations

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

$$\overline{\alpha}_{k+1} = \min \left\{ \sqrt{\frac{\sigma_k}{\sigma_{k-1}}\left(1+\rho_k\right)}\frac{\sigma_k}{\sigma_{k+1}}\alpha_k,\ \frac{\sqrt{1-4\left(1-\frac{\sigma_k}{\sigma_{k-1}}\right)\alpha_k \ell_k^{(2)}}}{2\sqrt{\alpha_k \ell_k [\alpha_k c_k - 1]_+}}\frac{\sigma_k}{\sigma_{k+1}}\alpha_k,\ \alpha_{\max} \right\}$$

- $\ell_k = \dfrac{\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1} \rangle}{\|x^k - x^{k-1}\|^2}$ $\quad \ell_k^{(2)} = \dfrac{\langle \nabla f^{(2)}(x^k) - \nabla f^{(2)}(x^{k-1}), x^k - x^{k-1} \rangle}{\|x^k - x^{k-1}\|^2}$
- $c_k = \dfrac{\|\nabla f_k(x^k) - \nabla f_k(x^{k-1})\|^2}{\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1} \rangle}$ $\quad \rho_k = \dfrac{\sigma_k \alpha_k}{\sigma_{k-1}\alpha_{k-1}}$

recovers adaPGM[1] (previous talk) when σ_k constant

[1] P. Latafat, AT, L. Stella and P. Patrinos, *"Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient"*, arXiv:2301.04431, 2023

An adaptive proximal gradient solver

adaBiM

adaBiM $(f^{(1)}, f^{(2)} \in C^{1,+})$

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

$$\overline{\alpha}_{k+1} = \min\left\{ \sqrt{\frac{\sigma_k}{\sigma_{k-1}}\left(1+\rho_k\right)}\,\frac{\sigma_k}{\sigma_{k+1}}\,\alpha_k,\ \frac{\sqrt{1-4\left(1-\frac{\sigma_k}{\sigma_{k-1}}\right)\alpha_k \ell_k^{(2)}}}{2\sqrt{\alpha_k \ell_k [\alpha_k c_k - 1]_+}}\,\frac{\sigma_k}{\sigma_{k+1}}\,\alpha_k,\ \alpha_{\max} \right\}$$

- $\ell_k = \dfrac{\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1}\rangle}{\|x^k - x^{k-1}\|^2}$ $\qquad$ $\ell_k^{(2)} = \dfrac{\langle \nabla f^{(2)}(x^k) - \nabla f^{(2)}(x^{k-1}), x^k - x^{k-1}\rangle}{\|x^k - x^{k-1}\|^2}$

- $c_k = \dfrac{\|\nabla f_k(x^k) - \nabla f_k(x^{k-1})\|^2}{\langle \nabla f_k(x^k) - \nabla f_k(x^{k-1}), x^k - x^{k-1}\rangle}$ $\qquad$ $\rho_k = \dfrac{\sigma_k \alpha_k}{\sigma_{k-1}\alpha_{k-1}}$

recovers adaPGM[1] (previous talk) when σ_k constant

✓ $\alpha_{\max}$ can be set arbitrarily large *(speed/backtracks unaffected)*

[1] P. Latafat, AT, L. Stella and P. Patrinos, *"Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient"*, arXiv:2301.04431, 2023

A nonadaptive proximal gradient solver

staBiM

staBiM ($f^{(1)}, f^{(2)} \in C^{1,1}$)

$$\begin{aligned} &\text{minimize } f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in \arg\min f^{(2)} + g^{(2)} \end{aligned}$$

When $\nabla f^{(i)}$ are *globally* Lipschitz with known moduli $L_{f^{(i)}}$,

$$\alpha_k = \frac{\nu}{\sigma_k L_{f^{(1)}} + L_{f^{(2)}}}$$

works, **without linesearch**

- ✗ Conservative choice, prone to slower convergence
- Even if $L_{f^{(i)}}$ known, *adaptive tuning* of adaBiM is preferable

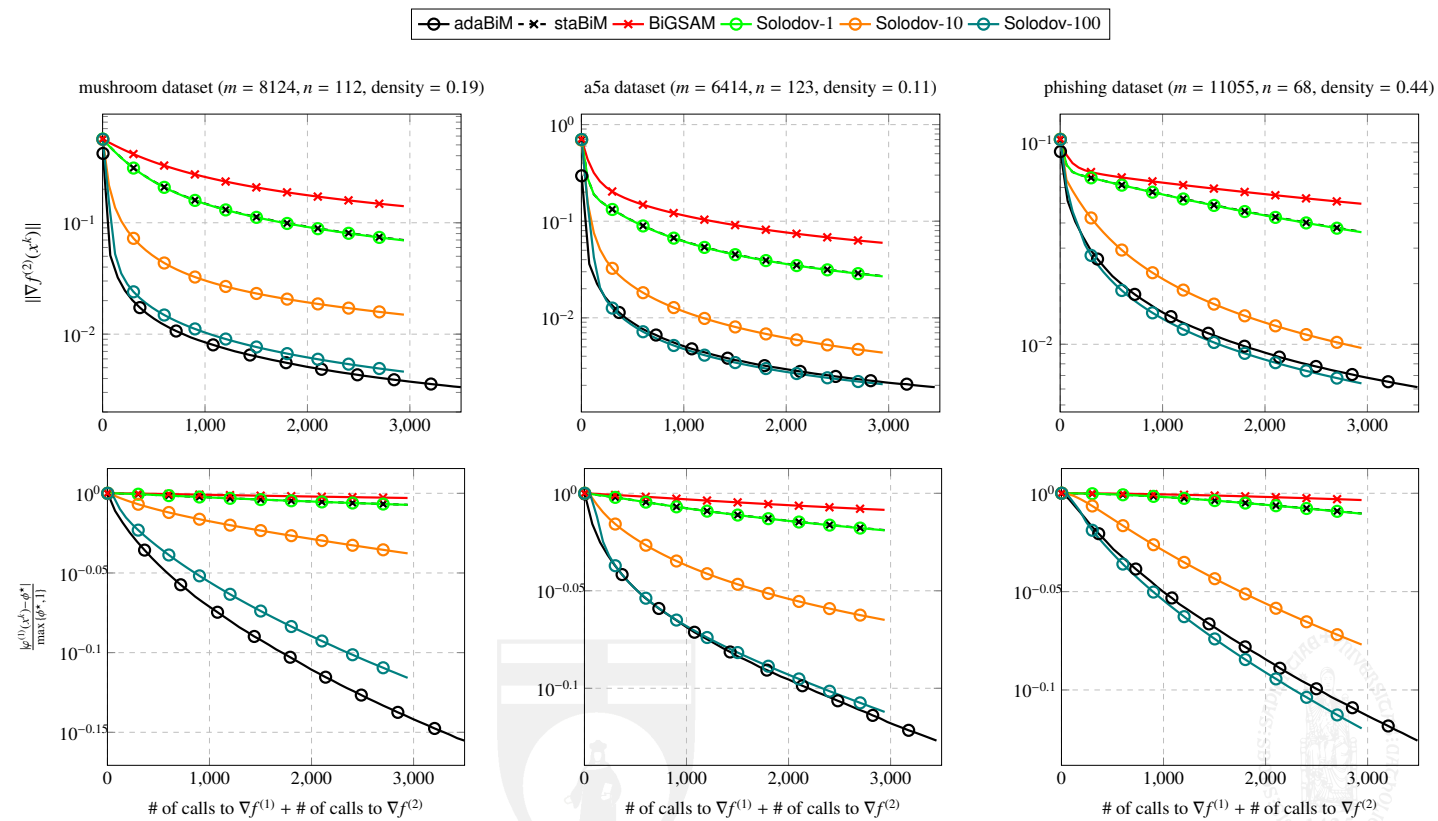
Outline

Simulations

Minimum ℓ^2-norm logistic regression

$$\begin{aligned} &\text{minimize } \|x\|^2 \\ &\text{subject to } x \in \arg\min_w \tfrac{1}{m} \textstyle\sum_{i=1}^m \{y_i \log s_i(w) + (1-y_i)\log(1-s_i(w))\} \end{aligned}$$

- $s_i(x) = (1+\exp(-a_i^\top x))^{-1}$ logistic sigmoid function
- in Solodov-c, $\alpha_{\max} = c/L_{f(1)}$



Simulations

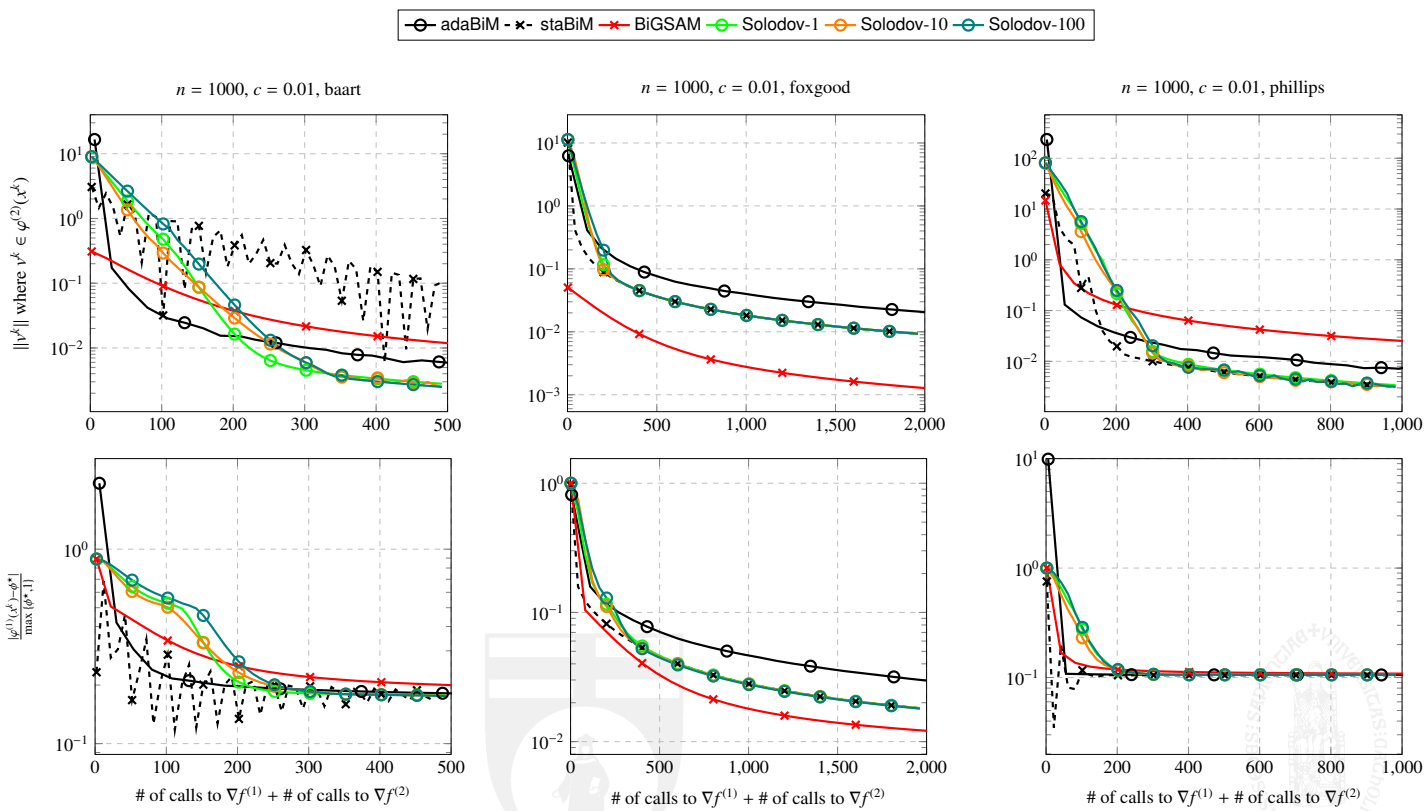
Solution of integral equations

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\|x\|_Q^2 \\ &\text{subject to } x \in \arg\min_w \left\{ \|Aw - b\|^2 + \delta_{\geqslant 0}(w) \right\} \end{aligned}$$

- $Q = L^\top L + \mathrm{I}$, with L discrete gradient
- in Solodov-c, $\alpha_{\max} = {}^{c}/_{L_{f^{(1)}}}$

feasibility
$\|\partial \varphi^{(2)}\|$

optimality
$\frac{|f^{(1)} - \phi^\star|}{\max\{1, |\phi^\star|\}}$



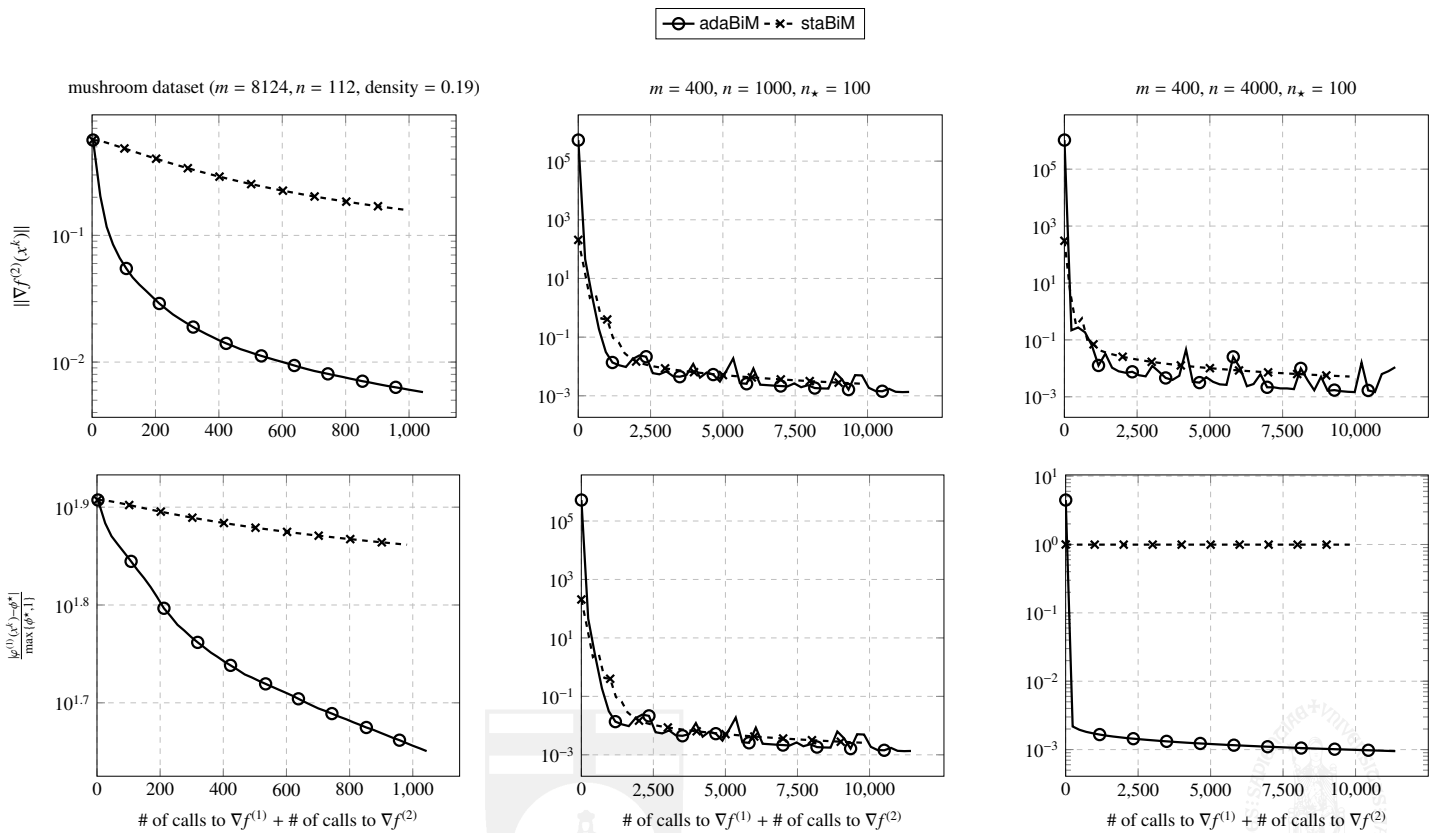
Simulations

Minimum ℓ^1-norm problems

- logistic regression and linear inverse problems ($\varphi^{(1)} = g^{(1)} = \|\cdot\|_1$)
- only adaBiM and staBiM can handle a nonsmooth upper cost
- staBiM can stagnate on ill-conditioned problems

feasibility
$\|\partial \varphi^{(2)}\|$

optimality
$\frac{|\varphi^{(1)} - \phi^\star|}{\max\{1, |\phi^\star|\}}$



Simulations

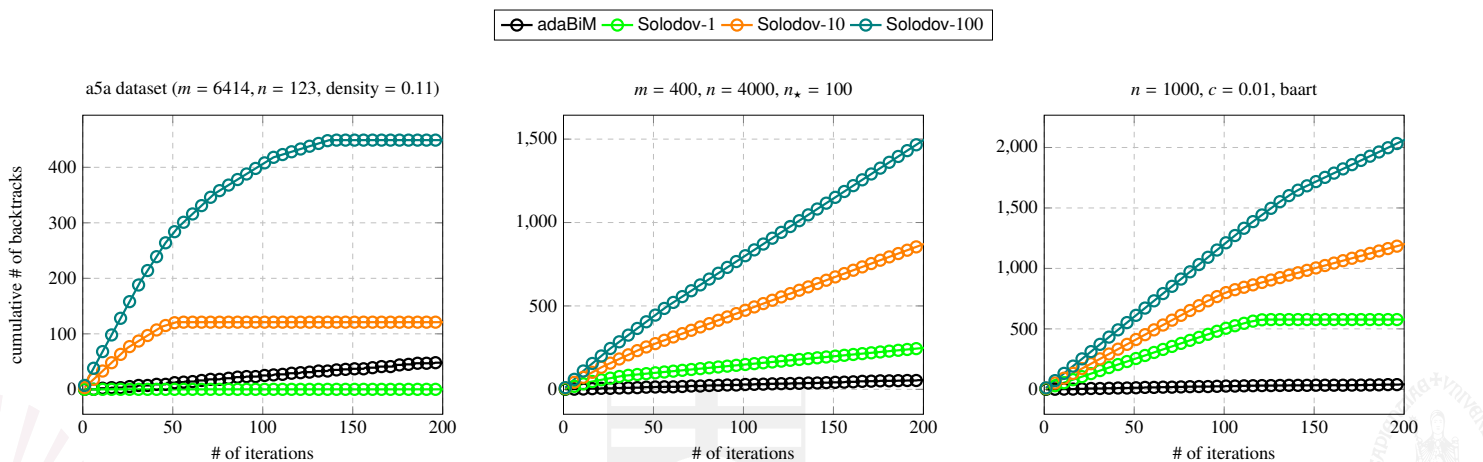
Cumulative number of backtracks

adaBiM

- unaffected by parameters choice
- # of backtracks counted in previous plots
  (each involves 1 ∇f_k-eval)

Solodov

- $\alpha_{\max}$ severly affects performance
- # of backtracks not counted in previous plots
  (each involves 1 f_k-eval)



Outline

Conclusions

Bilevel problems

- very general framework
- active area of research

adaBiM

- extends adaPG[2] to bilevel problems
- handles nonsmoothness on both levels
- ~parameter-free, (*fully* adaptive)
- linesearch **not wasteful** (even when not needed)

All the details in

> P. Latafat, AT, S. Villa and P. Patrinos, *"AdaBiM: An adaptive proximal gradient method for structured convex bilevel optimization"*, arXiv:2305.03559, 2023

Open problems

- ? termination criteria
- ? non"simple" bilevel programs
- ? σ_k: adaptive penalty selection

[2] P. Latafat, AT, L. Stella and P. Patrinos, *"Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient"*, arXiv:2301.04431, 2023