A Conditional Gradient Method for Simple Bilevel Optimization with Convex Lower-level Problem

Aryan Mokhtari
ECE Department, UT Austin

Joint work with
Ruichen Jiang (UT Austin), Nazanin Abolfazli (U Arizona),
and Erfan Y. Hamedani (U Arizona)

SIAM Conference on Optimization 2023, Seattle

Bilevel Optimization: General Form

- Bilevel optimization is a form of optimization where one problem is embedded within another.

General form of Bilevel Optimization Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{w}\in\mathbb{R}^m} \; f(\mathbf{x},\mathbf{w}) \qquad \textbf{s.t.} \quad \mathbf{x}\in\arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z},\mathbf{w}). \tag{GBO}$$

- In the general case that both f and g are only convex:
  - GBO is NP-hard [Vicente et al.'94]

Bilevel Optimization: General Form

- Bilevel optimization is a form of optimization where one problem is embedded within another.

General form of Bilevel Optimization Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{w}\in\mathbb{R}^m} \; f(\mathbf{x},\mathbf{w}) \qquad \textbf{s.t.} \;\; \mathbf{x}\in\arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z},\mathbf{w}). \tag{GBO}$$

- In the general case that both f and g are only convex:
  - GBO is NP-hard [Vicente et al.'94]

- To address this issue, two different settings are considered in the literature
  - Assuming that the lower-level problem is strongly-convex wrt to $\mathbf{z}$
  - Studying a simpler version of GBO known as Simple Bilevel Optimization

    $\Rightarrow$ The focus of this talk!

Bilevel Optimization: General Form

- GBO can also be written as:

General form of Bilevel Optimization Problem:

$$\min_{\mathbf{w}} \; l(\mathbf{w}) := f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}), \qquad \textbf{where} \;\; \mathbf{x}^*(\mathbf{w}) \in \arg\min_{\mathbf{z} \in \mathcal{Z}} \; g(\mathbf{z}, \mathbf{w})$$

In this case we have:

$$\nabla \ell(\mathbf{w}) = \nabla_w f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}) - \nabla_{xw} g(\mathbf{y}^*(\mathbf{w}), \mathbf{w}) [\nabla_{xx} g(\mathbf{x}^*(\mathbf{w}), \mathbf{w})]^{-1} \nabla_x f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}).$$

Bilevel Optimization: General Form

- GBO can also be written as:

General form of Bilevel Optimization Problem:

$$\min_{\mathbf{w}} \; l(\mathbf{w}) := f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}), \qquad \textbf{where} \;\; \mathbf{x}^*(\mathbf{w}) \in \arg\min_{\mathbf{z} \in \mathcal{Z}} \; g(\mathbf{z}, \mathbf{w})$$

In this case we have:

$$\nabla \ell(\mathbf{w}) = \nabla_w f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}) - \nabla_{xw} g(\mathbf{y}^*(\mathbf{w}), \mathbf{w})[\nabla_{xx} g(\mathbf{x}^*(\mathbf{w}), \mathbf{w})]^{-1} \nabla_x f(\mathbf{x}^*(\mathbf{w}), \mathbf{w}).$$

- Under strong convexity of g, the above expression is well-defined and one can find an approximate stationary point of the loss ℓ
- Several works for this setting:
  - Implicit differentiation: [Domke,'12], [Pedregosa,'16] [Gould et al.,'16],[Ji et al.,'21], ...
  - Iterative differentiation: [Maclaurin et al.,'15], [Franceschi et al.,'18], ...
  - .....

Simple Bilevel Optimization

- The alternative approach for a computationally tractable case
  $\Rightarrow$ Eliminating the variable $\mathbf{w}$ $\Rightarrow$ Simple Bilevel Optimization

Simple bilevel optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \textbf{s.t.} \quad \mathbf{x} \in \operatorname*{arg\,min}_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}), \tag{SBO}$$

Simple Bilevel Optimization

- The alternative approach for a computationally tractable case
  $\Rightarrow$ Eliminating the variable $\mathbf{w}$ $\Rightarrow$ Simple Bilevel Optimization

Simple bilevel optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \textbf{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z}), \tag{SBO}$$

- $f, g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable functions on an open set containing $\mathcal{Z}$.
- g is convex, but not necessarily strongly convex.
  - Indeed, in this setting, ideas from the previous slide do not work!
- $\mathcal{Z}$ is a compact convex set.

Motivating Examples for SBO

- The following general form:

$$\min_{\mathbf{x}} \underbrace{\text{model validation loss}}_{\text{secondary objective}} \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z} \in \mathcal{Z}} \underbrace{\text{model training loss}}_{\text{primal objective}}$$

  - Over-parameterized regression: [Gao et al.'22]

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{A}_{\text{val}}\boldsymbol{\beta} - \mathbf{b}_{\text{val}}\|_2^2 \qquad \text{s.t.} \quad \boldsymbol{\beta} \in \arg\min_{\|\mathbf{z}\|_1 \leq \lambda} \frac{1}{2}\|\mathbf{A}_{\text{tr}}\mathbf{z} - \mathbf{b}_{\text{tr}}\|_2^2.$$

Motivating Examples for SBO

- The following general form:

$$\min_{\mathbf{x}} \underbrace{\text{model validation loss}}_{\text{secondary objective}} \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \underbrace{\text{model training loss}}_{\text{primal objective}}$$

  - Over-parameterized regression: [Gao et al.'22]

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{A}_{\text{val}}\boldsymbol{\beta} - \mathbf{b}_{\text{val}}\|_2^2 \qquad \text{s.t.} \quad \boldsymbol{\beta} \in \arg\min_{\|\mathbf{z}\|_1\leq\lambda} \frac{1}{2}\|\mathbf{A}_{\text{tr}}\mathbf{z} - \mathbf{b}_{\text{tr}}\|_2^2.$$

- Lifelong learning or continual learning
  - Continual dictionary learning

$$\min_{\tilde{\mathbf{D}}\in\mathbb{R}^{m\times q}} \min_{\tilde{\mathbf{X}}\in\mathbb{R}^{q\times n'}} \underbrace{\frac{1}{2n'}\sum_{k=1}^{n'}\|\mathbf{a}_k' - \tilde{\mathbf{D}}\tilde{\mathbf{x}}_k\|_2^2}_{\text{Error on new dataset}}$$

$$\text{s.t.} \quad \|\tilde{\mathbf{x}}_k\|_1 \leq \delta, k = 1,\ldots,n'; \ \tilde{\mathbf{D}} \in \arg\min_{\|\tilde{\mathbf{d}}_j\|_2\leq 1} \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{a}_i - \tilde{\mathbf{D}}\hat{\mathbf{x}}_i\|_2^2}_{\text{Error on old dataset}}$$

Related Work

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}), \tag{SBO}$$

1) Primal-Dual Algorithms:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{Z},\, g(\mathbf{x}) \leq g^*,$$

Related Work

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z}), \tag{SBO}$$

1) Primal-Dual Algorithms:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{Z},\, g(\mathbf{x}) \leq g^*,$$

- Strict feasibility and slater's condition may not hold
- What if we relax it and use $g(\mathbf{x}) \leq g^* + \epsilon$?
  $\Rightarrow$ norm of the optimal dual variable becomes very large for small ϵ
  $\Rightarrow$ causes instability and the upper bound on error could blow up

Related Work

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}), \tag{SBO}$$

1) Primal-Dual Algorithms:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{Z},\, g(\mathbf{x}) \leq g^*,$$

- Strict feasibility and slater's condition may not hold
- What if we relax it and use $g(\mathbf{x}) \leq g^* + \epsilon$?
  $\Rightarrow$ norm of the optimal dual variable becomes very large for small ϵ
  $\Rightarrow$ causes instability and the upper bound on error could blow up

2) Tikhonov-type regularization [Tikhonov-Arsenin'97]

Related Work

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \operatorname*{arg\,min}_{\mathbf{z}\in\mathcal{Z}} \ g(\mathbf{z}), \tag{SBO}$$

1) Primal-Dual Algorithms:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{Z},\, g(\mathbf{x}) \le g^*,$$

- Strict feasibility and slater's condition may not hold
- What if we relax it and use $g(\mathbf{x}) \le g^* + \epsilon$?
  $\Rightarrow$ norm of the optimal dual variable becomes very large for small ϵ
  $\Rightarrow$ causes instability and the upper bound on error could blow up

2) Tikhonov-type regularization [Tikhonov-Arsenin'97]
- Combining the two objective functions using a regularization parameter $\sigma > 0$

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \sigma g(\mathbf{x})$$

- Under certain assumptions the solution of (SBO) exactly matches with the regularized problem [Friedlander-Tseng'08], [Dempe et. al'21]
- Proposed adjusting the regularization parameter σ dynamically [Cabot'05], [Solodov'07] $\Rightarrow$ but all are asymptotic results.

(ϵ_f, ϵ_g)-optimal solution

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}),$$

Definition:

When f is convex, a point $\hat{\mathbf{x}} \in \mathcal{Z}$ is (ϵ_f, ϵ_g)-optimal for the bilevel problem in (SBO) if

$$f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f \quad \text{and} \quad g(\hat{\mathbf{x}}) - g^* \leq \epsilon_g.$$

When f is non-convex, $\hat{\mathbf{x}} \in \mathcal{Z}$ is (ϵ_f, ϵ_g)-optimal if

$$\mathcal{G}(\hat{\mathbf{x}}) \leq \epsilon_f \quad \text{and} \quad g(\hat{\mathbf{x}}) - g^* \leq \epsilon_g,$$

where $\mathcal{G}(\hat{\mathbf{x}})$ is the FW gap defined by

$$\mathcal{G}(\hat{\mathbf{x}}) \triangleq \max_{\mathbf{s}\in\mathcal{X}_g^*} \{\langle \nabla f(\hat{\mathbf{x}}), \hat{\mathbf{x}} - \mathbf{s}\rangle\}.$$

- where $g^* \triangleq \min_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z})$ and $\mathcal{X}_g^* \triangleq \arg\min_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z})$

Related work

References	Upper level	Lower level		Convergence		Oracle
	Objective f	Objective g	Feasible set $\mathcal{Z}$	Upper level	Lower level	
MNG [A. Beck& S. Sabach'14]	SC, differentiable	C, smooth	Closed	Asymptotic	$\mathcal{O}(1/\epsilon^2)$	projection
BiG-SAM [S. Sabach & S. Shtern'17]	SC, smooth	C, composite	Closed	Asymptotic	$\mathcal{O}(1/\epsilon)$	projection
Tseng's method [Y. Malitsky'17]	C, composite	C, composite	Closed	Asymptotic	$o(1/\epsilon)$	projection
a-IRG [H.D. Kaushik & F. Yousefian'21]	C, Lipschitz	VI, Lipschitz	Closed	$\mathcal{O}(\max\{1/\epsilon_f^4, 1/\epsilon_g^4\})$		projection
Ours	C, smooth	C, smooth	Compact	$\mathcal{O}(\max\{1/\epsilon_f, 1/\epsilon_g\})$		linear solver
Ours	Non-C, smooth	C, smooth	Compact	$\mathcal{O}(\max\{1/\epsilon_f^2, 1/(\epsilon_f\epsilon_g)\})$		linear solver

Challenges

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \qquad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{x}\in\mathcal{X}_g^*} f(\mathbf{x})$$

Challenges

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) \qquad \text{s.t.} \ \ \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \ g(\mathbf{z}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{x}\in\mathcal{X}_g^*} f(\mathbf{x})$$

Frank Wolfe Method:

- Find a feasible direction by minimizing linear approximation of objective

$$\mathbf{s}_k \leftarrow \arg\min_{\mathbf{s}\in\mathcal{X}_g^*} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$$

$$\mathbf{x}_{k+1} \leftarrow (1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$$

Challenges

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \qquad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{x}\in\mathcal{X}_g^*} f(\mathbf{x})$$

Frank Wolfe Method:

- Find a feasible direction by minimizing linear approximation of objective

$$\mathbf{s}_k \leftarrow \arg\min_{\mathbf{s}\in\mathcal{X}_g^*} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$$

$$\mathbf{x}_{k+1} \leftarrow (1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$$

- Challenge I: $\mathcal{X}_g^*$ for the lower-level problem is not explicitly given
  $\Rightarrow$ The linear minimization is computationally intractable

Challenges

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) \qquad \text{s.t.} \ \ \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \ g(\mathbf{z}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{x}\in\mathcal{X}_g^*} f(\mathbf{x})$$

Frank Wolfe Method:

- Find a feasible direction by minimizing linear approximation of objective

$$\mathbf{s}_k \leftarrow \arg\min_{\mathbf{s}\in\mathcal{X}_g^*} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$$

$$\mathbf{x}_{k+1} \leftarrow (1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$$

- Challenge I: $\mathcal{X}_g^*$ for the lower-level problem is not explicitly given
  $\Rightarrow$ The linear minimization is computationally intractable

- Challenge II: The FW method needs to be initialized with a feasible point
  $\Rightarrow$ requires exact solution of g that is computationally intractable

Challenges

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) \qquad \text{s.t.} \;\; \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathcal{Z}} \; g(\mathbf{z}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{x}\in\mathcal{X}_g^*} f(\mathbf{x})$$

Frank Wolfe Method:

- Find a feasible direction by minimizing linear approximation of objective

$$\mathbf{s}_k \leftarrow \arg\min_{\mathbf{s}\in\mathcal{X}_g^*} \langle \nabla f(\mathbf{x}_k), \mathbf{s}\rangle$$

$$\mathbf{x}_{k+1} \leftarrow (1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$$

- Challenge I: $\mathcal{X}_g^*$ for the lower-level problem is not explicitly given
  $\Rightarrow$ The linear minimization is computationally intractable
- Challenge II: The FW method needs to be initialized with a feasible point
  $\Rightarrow$ requires exact solution of g that is computationally intractable
- Remark: Similar issues also hold for projection-based methods

CG-BiO: Main Idea

- Resolution: Provide an approximation for $\mathcal{X}_g^*$

CG-BiO: Main Idea

- Resolution: Provide an approximation for $\mathcal{X}_g^*$
- What properties should the approximation set $\mathcal{X}_k$ have?
  - It should have an explicit expression so that LMO is tractable
  - It should contain the set $\mathcal{X}_g^*$
  - It should allow us to control the increase of g

CG-BiO: Main Idea

- Resolution: Provide an approximation for $\mathcal{X}_g^*$
- What properties should the approximation set $\mathcal{X}_k$ have?
  - It should have an explicit expression so that LMO is tractable
  - It should contain the set $\mathcal{X}_g^*$
  - It should allow us to control the increase of g

Our idea:

- Consider $\mathbf{x}_0$ as an $\frac{\epsilon_g}{2}$ approximate solution of the lower-level problem

  $\Rightarrow \mathbf{x}_0 \in \mathcal{Z}$ and $g(\mathbf{x}_0) - g^* \leq \frac{\epsilon_g}{2}$ $\Rightarrow$ it is easy to find such a point

CG-BiO: Main Idea

- Resolution: Provide an approximation for $\mathcal{X}_g^*$
- What properties should the approximation set $\mathcal{X}_k$ have?
  - It should have an explicit expression so that LMO is tractable
  - It should contain the set $\mathcal{X}_g^*$
  - It should allow us to control the increase of g

Our idea:

- Consider $\mathbf{x}_0$ as an $\frac{\epsilon_g}{2}$ approximate solution of the lower-level problem

  $\Rightarrow \mathbf{x}_0 \in \mathcal{Z}$ and $g(\mathbf{x}_0) - g^* \leq \frac{\epsilon_g}{2}$ $\Rightarrow$ it is easy to find such a point

- We use the following set:

  $\mathcal{X}_k \triangleq \mathcal{Z} \cap \mathcal{H}_k$ where $\mathcal{H}_k = \{\mathbf{s} \in \mathbb{R}^n : \langle \nabla g(\mathbf{x}_k), \mathbf{s} - \mathbf{x}_k \rangle \leq g(\mathbf{x}_0) - g(\mathbf{x}_k)\}$

CG-BiO: Main Idea

- Resolution: Provide an approximation for $\mathcal{X}_g^*$
- What properties should the approximation set $\mathcal{X}_k$ have?
  - It should have an explicit expression so that LMO is tractable
  - It should contain the set $\mathcal{X}_g^*$
  - It should allow us to control the increase of g

Our idea:

- Consider $\mathbf{x}_0$ as an $\frac{\epsilon_g}{2}$ approximate solution of the lower-level problem

  $\Rightarrow \mathbf{x}_0 \in \mathcal{Z}$ and $g(\mathbf{x}_0) - g^* \leq \frac{\epsilon_g}{2}$ $\Rightarrow$ it is easy to find such a point

- We use the following set:

$$\mathcal{X}_k \triangleq \mathcal{Z} \cap \mathcal{H}_k \quad \text{where } \mathcal{H}_k = \{\mathbf{s} \in \mathbb{R}^n : \langle \nabla g(\mathbf{x}_k), \mathbf{s} - \mathbf{x}_k \rangle \leq g(\mathbf{x}_0) - g(\mathbf{x}_k)\}$$

- Our Method: At each iteration follow

$$\mathbf{x}_{k+1} \leftarrow (1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k, \qquad \text{where} \quad \mathbf{s}_k \leftarrow \arg\min_{\mathbf{s} \in \mathcal{X}_k} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$$

Why does it make sense?

It checks all the boxes!

Why does it make sense?

It checks all the boxes!

- The cutting plane $\mathcal{H}_k$ and the approximation set $\mathcal{X}_k$ are explicit

Why does it make sense?

It checks all the boxes!

- The cutting plane $\mathcal{H}_k$ and the approximation set $\mathcal{X}_k$ are explicit
- $\mathcal{X}_g^* \subseteq \mathcal{X}_k$ for all $k \geq 0$. (Why?) if $\hat{\mathbf{x}} \in \mathcal{X}_g^*$, then
  - $\hat{\mathbf{x}} \in \mathcal{Z}$ and $\langle \nabla g(\mathbf{x}_k), \hat{\mathbf{x}} - \mathbf{x}_k \rangle \leq g^* - g(\mathbf{x}_k) \leq g(\mathbf{x}_0) - g(\mathbf{x}_k) \Rightarrow \hat{\mathbf{x}} \in \mathcal{X}_k$

Why does it make sense?

It checks all the boxes!

- The cutting plane $\mathcal{H}_k$ and the approximation set $\mathcal{X}_k$ are explicit
- $\mathcal{X}_g^* \subseteq \mathcal{X}_k$ for all $k \geq 0$. (Why?) if $\hat{\mathbf{x}} \in \mathcal{X}_g^*$, then
  - $\hat{\mathbf{x}} \in \mathcal{Z}$ and $\langle \nabla g(\mathbf{x}_k), \hat{\mathbf{x}} - \mathbf{x}_k \rangle \leq g^* - g(\mathbf{x}_k) \leq g(\mathbf{x}_0) - g(\mathbf{x}_k) \Rightarrow \hat{\mathbf{x}} \in \mathcal{X}_k$
- We can control the possible increase in g. (How?)

  By smoothness and construction of $\mathcal{X}_k$:

$$\begin{aligned} g(\mathbf{x}_{k+1}) &\leq g(\mathbf{x}_k) + \gamma_k \langle \nabla g(\mathbf{x}_k), \mathbf{s}_k - \mathbf{x}_k \rangle + \frac{L_g \gamma_k^2 D^2}{2} \\ &\leq g(\mathbf{x}_k) + \gamma_k (g(\mathbf{x}_0) - g(\mathbf{x}_k)) + \frac{L_g \gamma_k^2 D^2}{2} \end{aligned}$$

  Hence,

$$g(\mathbf{x}_{k+1}) - g(\mathbf{x}_0) \leq (1 - \gamma_k)(g(\mathbf{x}_k) - g(\mathbf{x}_0)) + \frac{L_g \gamma_k^2 D^2}{2}$$

CG-BiO Algorithm

Input: Target accuracy $\epsilon_f, \epsilon_g > 0$, stepsizes $\{\gamma_k\}_k$
Initialization: Initialize $\mathbf{x}_0 \in \mathcal{Z}$ such that $0 \leq g(\mathbf{x}_0) - g^* \leq \epsilon_g/2$
for $k = 0, \ldots, K-1$ **do**
    Compute $\mathbf{s}_k \leftarrow \arg\min_{\mathbf{s} \in \mathcal{X}_k} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$
    if $\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}_k \rangle \leq \epsilon_f$ and $\langle \nabla g(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}_k \rangle \leq \epsilon_g/2$ **then**
        Return $\mathbf{x}_k$ and STOP
    else
        $\mathbf{x}_{k+1} \leftarrow (1 - \gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$
    end if
end for

Convergence Analysis in Convex Setting

Theorem 1 (Convex upper-level)

Let $\{\mathbf{x}_k\}_{k=0}^{K}$ be the sequence generated by CG-BiO Algorithm with stepsize $\gamma_k = 2/(k+2)$ for $k \geq 0$. Then we have

$$f(\mathbf{x}_K) - f^* \leq \frac{2L_f D^2}{K+1} \quad \text{and} \quad g(\mathbf{x}_K) - g^* \leq \frac{2L_g D^2}{K+1} + \frac{1}{2}\epsilon_g.$$

Convergence Analysis in Convex Setting

Theorem 1 (Convex upper-level)

Let $\{\mathbf{x}_k\}_{k=0}^{K}$ be the sequence generated by CG-BiO Algorithm with stepsize $\gamma_k = 2/(k+2)$ for $k \geq 0$. Then we have

$$f(\mathbf{x}_K) - f^* \leq \frac{2L_f D^2}{K+1} \quad \text{and} \quad g(\mathbf{x}_K) - g^* \leq \frac{2L_g D^2}{K+1} + \frac{1}{2}\epsilon_g.$$

Algorithm CG-BiO will return an (ϵ_f, ϵ_g)-optimal solution when the number of iterations K exceeds

$$\max\left\{\frac{2L_f D^2}{\epsilon_f} - 1, \frac{4L_g D^2}{\epsilon_g} - 1\right\} = \mathcal{O}\left(\max\left\{\frac{1}{\epsilon_f}, \frac{1}{\epsilon_g}\right\}\right).$$

Convergence Analysis in Non-convex Setting

Theorem 2 (Non-Convex upper-level)

Let $\{\mathbf{x}_k\}_{k=0}^{K-1}$ be the sequence generated by CG-BiO Algorithm with step-size $\gamma_k = \min\left\{\frac{\epsilon_f}{L_f D^2}, \frac{\epsilon_g}{L_g D^2}\right\}$ for all $k \geq 0$. Define $\underline{f} = \min_{\mathbf{x} \in Z} f(\mathbf{x})$. Then for $K \geq \max\left\{\frac{2L_f D^2 (f(\mathbf{x}_0) - \underline{f})}{\epsilon_f^2}, \frac{2L_g D^2 (f(\mathbf{x}_0) - \underline{f})}{\epsilon_f \epsilon_g}\right\}$, there exists $k^* \in \{0, 1, \ldots, K-1\}$ such that $\mathcal{G}(\mathbf{x}_{k^*}) \leq \epsilon_f$ and $g(\mathbf{x}_{k^*}) - g^* \leq \epsilon_g$.

Convergence Analysis in Non-convex Setting

Theorem 2 (Non-Convex upper-level)

Let $\{\mathbf{x}_k\}_{k=0}^{K-1}$ be the sequence generated by CG-BiO Algorithm with step-size $\gamma_k = \min\left\{\frac{\epsilon_f}{L_f D^2}, \frac{\epsilon_g}{L_g D^2}\right\}$ for all $k \geq 0$. Define $\underline{f} = \min_{\mathbf{x} \in Z} f(\mathbf{x})$. Then for $K \geq \max\left\{\frac{2L_f D^2 (f(\mathbf{x}_0) - \underline{f})}{\epsilon_f^2}, \frac{2L_g D^2 (f(\mathbf{x}_0) - \underline{f})}{\epsilon_f \epsilon_g}\right\}$, there exists $k^* \in \{0, 1, \ldots, K-1\}$ such that $\mathcal{G}(\mathbf{x}_{k^*}) \leq \epsilon_f$ and $g(\mathbf{x}_{k^*}) - g^* \leq \epsilon_g$.

The number of iterations required to find an (ϵ_f, ϵ_g)-optimal solution is

$$\mathcal{O}\left(\max\left\{\frac{1}{\epsilon_f^2}, \frac{1}{(\epsilon_f \epsilon_g)}\right\}\right)$$

.

Hölderian Error Bound

- Presented bounds can only guarantee that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f$

Hölderian Error Bound

- Presented bounds can only guarantee that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f$
- But, ideally we would like a bound of the form $|f(\hat{\mathbf{x}}) - f^*| \leq \epsilon_f$

Hölderian Error Bound

- Presented bounds can only guarantee that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f$
- But, ideally we would like a bound of the form $|f(\hat{\mathbf{x}}) - f^*| \leq \epsilon_f$
- However, there has been shown a negative result in [Chen et al. '23]!
  - for any first-order method and a given number of iterations K, there exists an instance of SBO where $|f(\mathbf{x}_k) - f^*| > 1$ for all k.

Hölderian Error Bound

- Presented bounds can only guarantee that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f$
- But, ideally we would like a bound of the form $|f(\hat{\mathbf{x}}) - f^*| \leq \epsilon_f$
- However, there has been shown a negative result in [Chen et al. '23]!
  - for any first-order method and a given number of iterations K, there exists an instance of SBO where $|f(\mathbf{x}_k) - f^*| > 1$ for all k.
- Hence, we need a condition to control how much $\mathbf{x}$ and then $f(\mathbf{x})$ can change when $g(\mathbf{x})$ is slightly worse than g^*

Hölderian Error Bound

- Presented bounds can only guarantee that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon_f$
- But, ideally we would like a bound of the form $|f(\hat{\mathbf{x}}) - f^*| \leq \epsilon_f$
- However, there has been shown a negative result in [Chen et al. '23]!
  - for any first-order method and a given number of iterations K, there exists an instance of SBO where $|f(\mathbf{x}_k) - f^*| > 1$ for all k.
- Hence, we need a condition to control how much $\mathbf{x}$ and then $f(\mathbf{x})$ can change when $g(\mathbf{x})$ is slightly worse than g^*

Assumption 2

The function g satisfies the Hölderian error bound for some $\alpha > 0$ and $r \geq 1$,

$$\frac{\alpha}{r}\, \mathbf{dist}(\mathbf{x}, \mathcal{X}_g^*)^r \leq g(\mathbf{x}) - g^*, \qquad \forall \mathbf{x} \in \mathcal{Z},$$

- The Hölderian error bound holds generically for real analytic and subanalytic functions [Łojasiewicz'63], [Łojasiewicz'93]

Convergence Under Hölderian Error Bound

Proposition 1

If g satisfies the Hölderian error bound, and $M = \max_{\mathbf{x} \in \mathcal{X}_g^*} \|\nabla f(\mathbf{x})\|_*$, then for any $\hat{\mathbf{x}}$ that satisfies $g(\hat{\mathbf{x}}) - g^* \le \epsilon_g$, it holds that:

(i) If f is convex, then $f(\hat{\mathbf{x}}) - f^* \ge -M \left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{1}{r}}$.

(ii) If f is non-convex and has L_f-Lipschitz gradient, then $\mathcal{G}(\hat{\mathbf{x}}) \ge -M \left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{1}{r}} - L_f \left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{2}{r}}$.

Convergence Under Hölderian Error Bound

Proposition 1

If g satisfies the Hölderian error bound, and $M = \max_{\mathbf{x}\in\mathcal{X}_g^*} \|\nabla f(\mathbf{x})\|_*$, then for any $\hat{\mathbf{x}}$ that satisfies $g(\hat{\mathbf{x}}) - g^* \le \epsilon_g$, it holds that:

(i) If f is convex, then $f(\hat{\mathbf{x}}) - f^* \ge -M\left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{1}{r}}$.

(ii) If f is non-convex and has L_f-Lipschitz gradient, then $\mathcal{G}(\hat{\mathbf{x}}) \ge -M\left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{1}{r}} - L_f\left(\frac{r\epsilon_g}{\alpha}\right)^{\frac{2}{r}}$.

Corollary

Let g satisfies the Hölderian error bound

(i) If f in Problem (SBO) is convex, we can set $\epsilon_g = \mathcal{O}(\epsilon_f^r)$, then after $K = \mathcal{O}(1/\epsilon_f^r)$ iterations, $|f(\mathbf{x}_K) - f^*| \le \epsilon_f$ and $g(\mathbf{x}_K) - g^* \le \epsilon_g$.

(ii) If f in Problem (SBO) is non-convex, and $\epsilon_g = \mathcal{O}(\epsilon_f^r)$ Then after $K = \mathcal{O}(1/\epsilon_f^{r+1})$ iterations, there exists $k^* \in \{0, 1, \ldots, K-1\}$ such that $|\mathcal{G}(\mathbf{x}_{k^*})| \le \epsilon_f$ and $g(\mathbf{x}_{k^*}) - g^* \le \epsilon_g$.

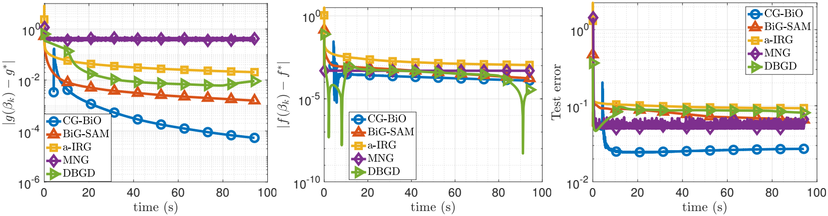
Numerical Experiments

Over-parameterized regression

- Sparse linear regression problem on the Wikipedia Math Essential dataset
- Data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and outcome vector $\mathbf{b} \in \mathbb{R}^{n}$, with $n = 1068$ instances and $d = 730$ attributes
- 60% as training set $(\mathbf{A}_{\rm tr}, \mathbf{b}_{\rm tr})$, 20% as validation set $(\mathbf{A}_{\rm val}, \mathbf{b}_{\rm val})$, the rest as test set
- The bilevel formulation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \ f(\boldsymbol{\beta}) \triangleq \frac{1}{2}\|\mathbf{A}_{\rm val}\boldsymbol{\beta} - \mathbf{b}_{\rm val}\|_2^2$$

$$\text{s.t.} \quad \boldsymbol{\beta} \in \mathop{\arg\min}_{\|\mathbf{z}\|_1\le\lambda} \ g(\mathbf{z}) \triangleq \frac{1}{2}\|\mathbf{A}_{\rm tr}\mathbf{z} - \mathbf{b}_{\rm tr}\|_2^2.$$



Dictionary learning

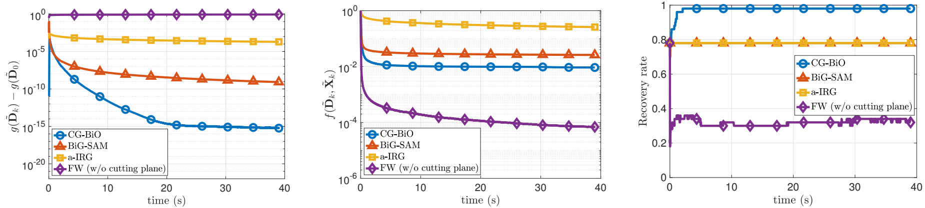
- Generate the true dictionary $\tilde{\mathbf{D}}^* \in \mathbb{R}^{25\times 50}$.
- Construct the two dictionaries $\mathbf{D}^*$ and $\mathbf{D}^{*\prime}$.
- Generate the two dataset $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{250}\}$ and $\mathbf{A}' = \{\mathbf{a}'_1, \ldots, \mathbf{a}'_{200}\}$ according to the following rules:

$$\mathbf{a}_i = \mathbf{D}^*\mathbf{x}_i + \mathbf{n}_i,\ i = 1, 2, \ldots, 250, \quad \mathbf{a}'_k \quad = \mathbf{D}'^*\mathbf{x}'_k + \mathbf{n}'_k,\ k = 1, 2, \ldots, 200,$$

Dictionary learning

- Generate the true dictionary $\tilde{\mathbf{D}}^* \in \mathbb{R}^{25\times 50}$.
- Construct the two dictionaries $\mathbf{D}^*$ and $\mathbf{D}^{*\prime}$.
- Generate the two dataset $\mathbf{A} = \{\mathbf{a}_1, \dots, \mathbf{a}_{250}\}$ and $\mathbf{A}' = \{\mathbf{a}'_1, \dots, \mathbf{a}'_{200}\}$ according to the following rules:

$$\mathbf{a}_i = \mathbf{D}^*\mathbf{x}_i + \mathbf{n}_i,\ i = 1, 2, \dots, 250, \quad \mathbf{a}'_k \quad = \mathbf{D}'^*\mathbf{x}'_k + \mathbf{n}'_k,\ k = 1, 2, \dots, 200,$$



Thank you

Any questions?

Based on:

- R. Jiang, N. Abolfazli, A. Mokhtari, E. Y. Hamedani, “A Conditional Gradient-based Method for Simple Bilevel Optimization with Convex Lower-level Problem", *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Valencia, Spain, 2023.