Tsinghua University

Functionally Constrained Algorithm Solves Convex Simple Bilevel Problems

Huaqing Zhang* [1,2], Lesi Chen* [1,2], Jing Xu[1], Jingzhao Zhang[1,2,3]

[1]IIIS, Tsinghua University [2]Shanghai Qizhi Institute [3] Shanghai AI Lab

Equal Contributions

Presenter: Huaqing Zhang

Ⅰ. Introduction

- **Our problem setup [Simple Bilevel Optimization]:**

$$\min_{\mathbf{x}\in\mathcal{Z}} f(\mathbf{x}) \ \text{s.t.}\ \mathbf{x}\in\mathcal{X}_g^* = \arg\min_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z}) \qquad (1)$$

$\mathcal{X}_g^*$: minimizers of $g(x)$
g^*: minimal of $g(x)$
f^*: minimal of $f(x)$ over $\mathcal{X}_g^*$

  - Minimize the upper-level objective over the solution set of a lower-level problem.
  - $\mathcal{Z}$: feasible set; convex & compact with diameter D.
  - $\boldsymbol{f}$ and $\boldsymbol{g}$: upper-level and lower-level objective functions.
    - Assumption 1: f, g are convex and L_f, L_g-smooth functions.
    - Assumption 2: f, g are convex and C_f, C_g-Lipschitz continuous functions.
- **a hierarchical structure!** 👉 many applications in machine learning
  - *Lifelong learning, lexicographic optimization…*
- Challenge: $\mathcal{X}_g^*$ is not explicitly given.
  - Thus methods for constrained problems projected gradient method and Frank-Wolfe method are not applicable.

I. Introduction

- **Our contribution:**
  - **Fundamental Difficulty of Simple BiO problems**: Prove the intractability of any zero-respecting first-order methods to find absolute optimal solutions.
  - **Near-Optimal Methods**: Propose a novel method with near-optimal rates for finding weak optimal solutions in both nonsmooth and smooth Simple BiO problems.

- **Absolute optimal solution**: $|f(\hat{x}) - f^*| \leq \epsilon_f, g(\hat{x}) - g^* \leq \epsilon_g.$
- **Weak optimal solution**: $f(\hat{x}) - f^* \leq \epsilon_f, g(\hat{x}) - g^* \leq \epsilon_g.$

Ⅱ. Hardness result: absolute optimal solution is not obtainable

- Our result: It is generally **intractable** for any *zero-respecting first-order method* to absolute optimal solutions.

Theorem 4.1. *For any first-order algorithm $\mathcal{A}$ satisfying Assumption 3.4 that runs for T iterations and any initial point $\mathbf{x}_0$, there exists a $(1,1)$-smooth instance of Problem (1) such that the optimal solution $\mathbf{x}^*$ satisfies $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq 1$ and $|f(\mathbf{x}_0) - f^*| \geq \frac{1}{48}$. For the iterates $\{\mathbf{x}_k\}_{k=0}^T$ generated by $\mathcal{A}$, the following holds:*

$$f(\mathbf{x}_k) = f(\mathbf{x}_0), \quad \forall 1 \leq k \leq T.$$

Theorem 4.2. *For any first-order algorithm $\mathcal{A}$ satisfying Assumption 3.4 that runs for T iterations and any initial point $\mathbf{x}_0$, there exists a $(1,1)$-Lipschitz instance of Problem (1) and some adversarial subgradients $\{\partial f(\mathbf{x}_k), \partial g(\mathbf{x}_k)\}_{k=0}^{T-1}$ such that the optimal solution $\mathbf{x}^*$ satisfies $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq 1$ and $|f(\mathbf{x}_0) - f^*| \geq \frac{1}{4}$. For the iterates $\{\mathbf{x}_k\}_{k=0}^T$ generated by $\mathcal{A}$, the following holds*

$$f(\mathbf{x}_k) = f(\mathbf{x}_0), \quad \forall 1 \leq k \leq T.$$

zero-respecting first-order method : $\mathcal{A}$ generates test points $\{\mathbf{x}_t\}_{t\geq 0}$ with

$$\mathrm{supp}(\mathbf{x}_{t+1}) \subseteq \mathrm{supp}(\mathbf{x}_0) \cup \left(\bigcup_{0\leq s\leq t} \mathrm{supp}(\partial f(\mathbf{x}_s)) \cup \mathrm{supp}(\partial g(\mathbf{x}_s))\right)$$

Ⅱ. Hardness result: absolute optimal solution is not obtainable

- Proof idea: we need to construct a “hard case”.
  - Key concept: “first-order zero-chain” (Definition 3.1)
    - Applying zero-respecting first-order method to a first-order zero-chain with zero initialization: only one component of x_k becomes non-zero in each iteration.

Ⅲ. Near-optimal method for finding weak-optimal solutions

- Due to the intractability of obtaining absolute optimal solutions, we focus on proposing first-order methods for finding weak-optimal solutions: $f(\hat{x}) - f^* \leq \epsilon_f, g(\hat{x}) - g^* \leq \epsilon_g$.
- **Step1**: reformulate the original simple BiO problem to a functionally constrained problem.

$$\boxed{\min f(x), s.t. \tilde{g}(x) := g(x) - \hat{g}^* \leq 0 \qquad (2)}$$

where $\hat{g}^*$ is an approximation of the lower-level problem's optimal value g^*

- **Step2**: Reduce Problem (2) to finding the smallest root of an auxiliary function (3), whose function value is defined by a discrete minimax problem. Such reformulation is introduced in Nesterov's *Lectures on convex optimization*.

$$\boxed{\psi^*(t) := \min_{x \in \mathcal{Z}} \{\psi(t,x) := \max\{f(x) - t, \tilde{g}(x)\}\} \quad (3)}$$

Nesterov, Yurii. *Lectures on convex optimization*. Vol. 137. Berlin: Springer, 2018.

Ⅲ. Near-optimal method for finding weak-optimal solutions

- To solve the smallest root of $\psi^*(t)$, we adopt a **bisection procedure**, and uses a **first-order subroutine** $\mathcal{M}$ to estimate the function value of $\psi^*(t)$ for a given t.

Algorithm: Functionally Constrained Bilevel Optimizer (FC-BiO)

Require: desired accuracy ϵ, total number of iterations T, initial bounds ℓ, u, and first-order subroutine $\mathcal{M}$.

Set $N = \left\lceil \log_2 \frac{u-\ell}{\epsilon/2} \right\rceil, K = T/N$. Set $\bar{\mathbf{x}} = \mathbf{x}_0$.

for $k = 0, \cdots, N-1$ **do**

  Set $t = \frac{\ell+u}{2}$.

  Solve with the subroutine $\left(\hat{\mathbf{x}}_{(t)}, \hat{\psi}^*(t)\right) = \mathcal{M}(\bar{\mathbf{x}}, t, K)$. Set $\bar{\mathbf{x}} = \hat{\mathbf{x}}_{(t)}$.

  if $\hat{\psi}^*(t) \geq \frac{\epsilon}{2}$ **then** set $\ell = t$.

  else set $u = t$.

End for

Return $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{(u)}$ as the approximate solution.

Ⅲ. Near-optimal method for finding weak-optimal solutions

first-order subroutine $\mathcal{M}$: $\min_{x\in\mathcal{Z}}\{\psi(t,x) := \max\{f(x)-t, \tilde{g}(x)\}\}$

Lipschitz objectives:

Subgradient Method

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{Z}}\left(\mathbf{x}_k - \eta\partial_{\mathbf{x}}\psi(t,\mathbf{x}_k)\right)$$

smooth objectives:

Generalized Accelerated Gradient Method

$$\mathbf{x}_{k+1} = \arg\min_{x\in\mathcal{Z}}\ \max\Big\{f(y_k) + \langle\nabla f(y_k), x-y_k\rangle + \frac{L}{2}\|x-y_k\|_2^2 - t, \\ \tilde{g}(y_k) + \langle\nabla\tilde{g}(y_k), x-y_k\rangle + \frac{L}{2}\|x-y_k\|_2^2\Big\}$$

$\mathbf{x}_{\boldsymbol{k+1}}$ can be further written in the form of a projection. (Proposition 5.2)

- Convergence rate of our FC-BiO method (Theorem 5.3, 5.4):

  - Lipschitz case:

$$\tilde{O}\left(\max\left\{\frac{C_f^2}{\epsilon_f^2}, \frac{C_g^2}{\epsilon_g^2}\right\} D^2\right)$$

  - Smooth case:

$$\tilde{O}\left(\max\left\{\sqrt{\frac{L_f}{\epsilon_f}}, \sqrt{\frac{L_g}{\epsilon_g}}\right\} D\right)$$

Where D is the diameter of $\mathcal{Z}$ C_f, C_g, L_f, L_g are Lipschitz/smooth constants, and $\tilde{O}$ hides logarithmic terms

Near-optimal rate in both settings

IV. Numerical experiment

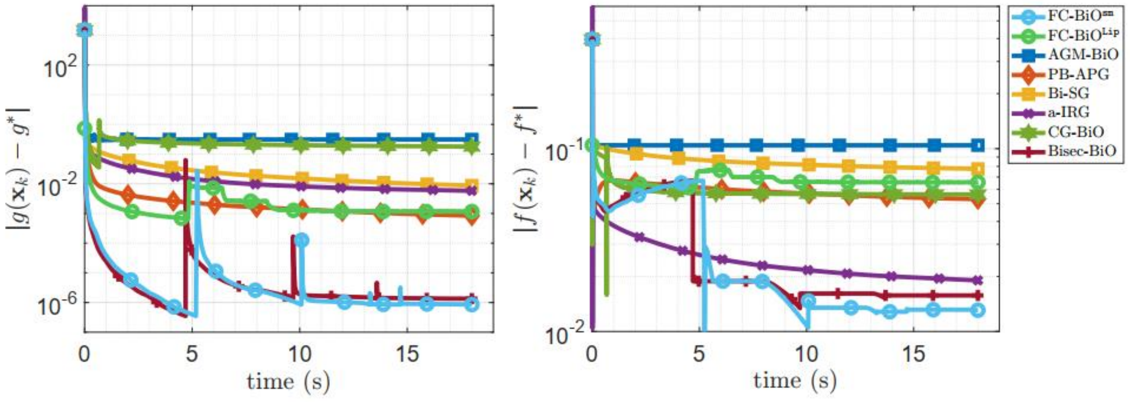

Figure 1: The performance of Algorithm 1 compared with other methods in Problem (11).

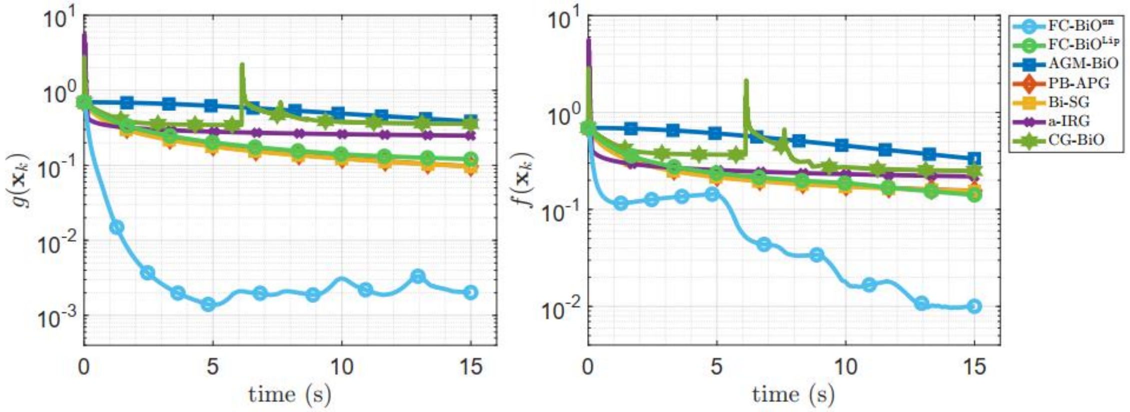

Figure 2: The performance of Algorithm 1 compared with other methods in Problem (12)

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2, \quad g(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - b\|_2^2,$$

minimum norm solution of Linear Regression.
400 datapoints. 700+ features.

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-(A_i^{val})^\top \mathbf{x}\mathbf{b}_i^{val})),$$
$$g(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-(A_i^{tr})^\top \mathbf{x}\mathbf{b}_i^{tr})).$$

Overparameterized Logistic Regression
10000 datapoints, 40000+ features.

Thank you!