On Bilevel Optimization without Lower-level Strong Convexity

Yifan Yang

Feb 22, 2023

Main Content

- Background
- Strong convexity of lower-level function
- Problem of non-strong convexity
- Continuity and Local Optimality of Hyper-Objective
- Goldstein Stationary Points and Algorithm
- Conclusions

Background

- What is bilevel optimization

$$\min_{x\in\mathbb{R}^d, y\in\mathcal{S}(x)} f(x,y), \quad \mathcal{S}(x) = \arg\min_{y\in\mathcal{Y}} g(x,y),$$

$$\varphi(x) \triangleq \min_{y\in\mathcal{S}(x)} f(x,y).$$

Background

- Bilevel vs minimax

Minimax: $$\min_x \max_y f(x, y)$$

Bilevel: $$\min_x f(x, y^*(x)), \quad y^*(x) = \arg\min_y - f(x, y)$$

Background

- Application

Supply chain management: Bilevel optimization can be used to model the interactions between suppliers and manufacturers. The lower level problem represents the production decisions of the manufacturer, while the upper level problem represents the pricing decisions of the supplier.

- Related Machine Learning Fields

Meta-learning, federated-learning, continual learning

Strong convexity of lower-level function

- Why we need that?

Strong convexity assumption on the lower-level objective makes the feasible set $S(x)$ a singleton, and the hyper-objective $\phi(x)$ a smooth, differentiable function.

$$\min_{x \in \mathbb{R}^d, y \in \mathcal{S}(x)} f(x, y), \quad \mathcal{S}(x) = \arg\min_{y \in \mathcal{Y}} g(x, y),$$

$$\varphi(x) \triangleq \min_{y \in \mathcal{S}(x)} f(x, y).$$

Solution with lower-level strong convexity

Implicit Differentiation Theorem.

For example, approximate implicit differentiation (AID) based methods and iterative differentiation (ITD) based methods.

$$\nabla \varphi(x) = \nabla_x f(x, y^*(x)) - \nabla_{xy}^2 g(x, y^*(x))[\nabla_{yy}^2 g(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)),$$

$$\text{where } y^*(x) \triangleq \arg\min_{y \in \mathbb{R}^q} g(x, y)$$

Problem of Non-strong Convexity

Discontinuity of hyper-objective

Example 3.1. *Consider the bilevel problem given by*

$$\min_{x \in \mathbb{R}, y \in \mathcal{S}(x)} (x^2 + 1)y, \quad S(x) = \arg\min_{y \in [-1,1]} -xy,$$

where the lower-level problem is convex in y. Then we know that

$$\varphi(x) = (x^2 + 1)\,\mathrm{sign}(x),$$

which is not continuous at the point $x = 0$.

Problem of Non-strong Convexity

The failure of Regularization

Proposition 3.1. *Consider regularizing the lower-level problem as* $g_\lambda(x,y) = g(x,y) + \frac{\lambda}{2}\|y - \hat{y}\|^2$ *for some* $\hat{y} \in \mathbb{R}^q$ *and* $\lambda > 0$*. There always exists a bilevel problem instance of form (1), where* $g(x,y)$ *is convex in* y, $\varphi(x)$ *is a quadratic function with stationary points, but* $\varphi_\lambda(x) = \min_{x \in \mathcal{S}_\lambda(x)} f(x,y), \mathcal{S}(x) = \arg\min_{y \in \mathcal{Y}} g_\lambda(x,y)$ *is a linear function without any stationary points. Additionally,* $|\min_x \varphi(x) - \min_{x'} \varphi_\lambda(x')| = \infty$.

$$\min_{x \in \mathbb{R}, y \in \mathcal{S}(x)} y_{[1]}^2 - 2xy_{[1]}, \quad \text{s.t. } S(x) = \arg\min_{y \in \mathbb{R}^2} (y_{[2]} - \hat{y}_{[2]})^2.$$

$$\varphi(x) = -x^2$$

$$\varphi_\lambda(x) = \hat{y}_{[1]}^2 - 2x\hat{y}_{[1]}$$

The feasible set will become a set to a unique point, so the original structure of the problem is broken.
And the **hyper-objective** and the **regulated hyper-objective** will completely different.

Problem of Non-strong Convexity

The failure of KKT condition

$$\min_{x\in\mathbb{R}^d, y\in\mathbb{R}^q} f(x,y), \quad \text{s.t. } g(x,y) \le \min_{y\in\mathcal{Y}} g(x,y),$$

There is no *(x, y)* such that the inequality holds.

The failure of Slater's condition!
Strong Duality doesn't hold!

$$\min_{x\in\mathbb{R}, y\in\mathbb{R}} -xy, \quad \text{s.t. } y \in \arg\min_{y\in\mathbb{R}} (x+y-2)^2,$$

Problem of Non-strong Convexity

Slater's condition can be satisfied for approximate KKT points.

$$\begin{cases} \lambda \geq 0; & \text{Non-negativity of multiplier} \\ g(x,y) - g^*(x) = \mathcal{O}(\varepsilon); & \text{Feasibility of constraint} \\ |\lambda(g(x,y) - g^*(x))| = \mathcal{O}(\varepsilon); & \text{Complementary slackness} \\ \mathrm{dist}(\nabla f(x,y) + \lambda(\nabla g(x,y) - \nabla g^*(x)), -\mathcal{N}(z;\mathcal{Z})) = \mathcal{O}(\varepsilon); & \text{Stationary of Lagrange function,} \end{cases}$$

But approximate KKT points can be problematic in the bilevel setting

$$\min_{x\in\mathbb{R}, y\in\mathbb{R}} x^2 - 2\varepsilon xy, \quad \text{s.t. } y \in \arg\min_{y\in\mathbb{R}} \varepsilon^3 y^2.$$

where the lower-level problem is strongly convex in y. There exists infinite $\mathcal{O}(\varepsilon)$-KKT points $(\hat{x}, \hat{y}, \hat{\lambda})$ such that $\|\nabla\varphi(\hat{x})\| = \Omega(1)$.

Continuity and Local Optimality of Hyper-Objective

Definition 4.1 (Hausdorff distance, [60]). *Let S_1, S_2 be two sets in $\mathbb{R}^d$. Define the Hausdorff distance of* $\mathrm{dist}(S_1, S_2)$ *by*

$$\mathrm{dist}(S_1, S_2) = \max\left\{ \sup_{x_1 \in S_1} \inf_{x_2 \in S_2} \|x_1 - x_2\|, \sup_{x_2 \in S_2} \inf_{x_1 \in S_1} \|x_1 - x_2\| \right\}$$

Based on the Hausdorff distance, we define the local Lipschitz continuity of the set-valued mapping $\mathcal{S}(x)$.

Definition 4.2. *We call a set-valued mapping $\mathcal{S}(x)$ locally Lipschitz continuous if for any $x \in \mathbb{R}^d$, there exists $\delta > 0$ and $L > 0$ such that for any $x' \in \mathbb{R}^d$ satisfying $\|x' - x\| \leq \delta$, we have* $\mathrm{dist}(\mathcal{S}(x), \mathcal{S}(x')) \leq L\|x - x'\|$.

As mentioned before, we assume the following nonemptiness and compactness of $\mathcal{S}(x)$ to ensure that the lower-level minimization is well defined.

Theorem:

If for any given *x* the set *S(x)* is non-empty and compact, and *f(x,y)* and *S(x)* are locally Lipschitz continuous, then *ϕ(x)* is locally Lipschitz continuous.

Continuity and Local Optimality of Hyper-Objective

Locally Lipschitz continuity of *S(x)*

Assumption 4.2 (Lipschitz objective with weak sharp minimum). *Suppose for any $x \in \mathbb{R}^d$ the lower-level problem g satisfies the following properties:*

- *Lipschitz in x for some constant $L > 0$:*

$$\|g(x,y) - g(x',y)\| \leq L\|x - x'\|, \quad \forall x, x' \in \mathbb{R}^d, y \in \mathcal{Y};$$

- *the optimal set of $g(x,\cdot)$ is the set of weak sharp minimum for some positive continuous function $\alpha(x)$:*

$$g(x,y) - \min_{y' \in \mathcal{Y}} g(x,y') \geq 2\alpha(x)\|y - y_p(x)\|, \quad \forall x \in \mathbb{R}^d, y \in \mathcal{Y},$$

where $y_p(x)$ is the projection of y onto the optimal set $\arg\min_{y' \in \mathcal{Y}} g(x,y')$.

Continuity and Local Optimality of Hyper-Objective

Locally Lipschitz continuity of *S(x)*

Assumption 4.3 (smooth objective with dominant gradient). *Suppose that the lower-level problem g satisfies the following properties:*

- *gradient Lipschitz for some $L > 0$:*

$$\|\nabla g(x, y) - \nabla g(x', y')\| \leq L(\|x - x'\| + \|y - y'\|), \quad \forall x, x' \in \mathbb{R}^d, y, y' \in \mathcal{Y};$$

- *gradient dominant in y for some positives continuous $\alpha(x)$:*

$$L \left\| y - \mathcal{P}_{\mathcal{Y}} \left(y - \frac{1}{L} \nabla_y g(x, y) \right) \right\| \geq \alpha(x) \|y - y_p(x)\|, \quad \forall x \in \mathbb{R}^d, y \in \mathcal{Y},$$

where $\mathcal{P}_{\mathcal{Y}}(\cdot)$ is the projection onto $\mathcal{Y}$ and $y_p(x)$ is the projection of y onto $\arg\min_{y \in \mathcal{Y}} g(x, y)$.

Continuity and Local Optimality of Hyper-Objective

Locally Lipschitz continuity of *S(x)*

Theorem:

Under Lipschitz objective with weak sharp minimum assumptions or smooth objective with dominant gradient assumptions, ***S(x)* is locally Lipschitz continuous**.

Furthermore, when $f(x,y)$ is locally Lipschitz continuous, then $\phi(x)$ is also locally Lipschitz continuous.

Previous Theorem:

If for any given x the set $S(x)$ is non-empty and compact, and $f(x,y)$ and $S(x)$ are locally Lipschitz continuous, then $\phi(x)$ is locally Lipschitz continuous.

Goldstein Stationary Points

Clarke subdifferential

$$\partial h(x) := \mathrm{Conv}\left\{ g : g = \lim_{x_k \to x} \nabla h(x_k) \right\}$$

(δ, ε)-Goldstein stationary point

$$\min\left\{ \|g\| : g \in \mathrm{Conv}\left\{ \cup_{x' \in \mathbb{B}_\delta(x)} \partial h(x') \right\} \right\} \leq \varepsilon,$$

A local minimum of *ϕ* must be a *(0, δ)*-Goldstein stationary point for any *δ* > 0.

Goldstein Stationary Points

Inexact Gradient-Free Method

Algorithm 1 IGFM

for $t = 0, 1, \cdots, T-1$
  Sample $u_t \in \mathbb{R}^d$ uniformly from the unit sphere in $\mathbb{R}^d$
  Estimate $\tilde{\varphi}(x_t + \delta u_t)$ and $\tilde{\varphi}(x_t - \delta u_t)$ by subroutine $\mathcal{A}$
  $\tilde{G}(x_t) = \frac{d}{2\delta}\left(\tilde{\varphi}(x_t + \delta u_t) - \tilde{\varphi}(x_t - \delta u_t)\right) u_t$
  $x_{t+1} = x_t - \eta \tilde{G}(x_t)$
end for
return $\bar{x}$ uniformly chosen from $\{x_t\}_{t=0}^{T-1}$

Goldstein Stationary Points

Inexact Gradient-Free Method

3: Estimate $\tilde{\varphi}(x_t + \delta u_t)$ and $\tilde{\varphi}(x_t - \delta u_t)$ by subroutine $\mathcal{A}$

4: $\tilde{G}(x_t) = \frac{d}{2\delta}\left(\tilde{\varphi}(x_t + \delta u_t) - \tilde{\varphi}(x_t - \delta u_t)\right) u_t$

Line 4 use zeroth-order method to calculate the hyper-gradient, but skip the design of estimation of ϕ.

Just an idea, not a solution. Because the subroutine *A* is the real difficulty in practice.

Conclusions

In this paper, the authors investigate the local optimality of bilevel optimization without lower-level strong convexity. They demonstrate that Goldstein stationary point can characterize the optimality for a general problem class, and propose the IGFM Algorithm for finding a Goldstein stationary point in polynomial time.

In my opinion, the highlights of this paper are those counterexamples which shows the limits of traditional non-strongly convex methods in bilevel condition.

References

Chen, Lesi and Xu, Jing and Zhang, Jingzhao. On Bilevel Optimization without Lower-level Strong Convexity

http://pi.math.cornell.edu/~web6630/Morse-Bott-talk.pdf

https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec13.pdf

https://people.math.wisc.edu/~ajnagel/convexity.pdf

https://arxiv.org/pdf/2002.04130.pdf