Theoretical smoothing frameworks for general nonsmooth bilevel problems

Jan Harold Alcantara
RIKEN AIP, Tokyo, Japan

International Symposium on Mathematical Programming (ISMP 2024)
Montreal, Canada

This is a joint work with Prof. Akiko Takeda (The University of Tokyo)

Outline

Bilevel optimization

We consider the general bilevel problem

$$\min_{(x,y)\in X\times Y} \quad f(x,y) \qquad \text{s.t.} \quad y \in \arg\min_{\bar{y}\in Y} g(x,\bar{y}) \tag{BP}$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ are possibly **nonsmooth** but **locally Lipschitz continuous**
- $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ are closed convex sets

Traditional approaches

- Replace the lower-level (LL) problem by its optimality conditions.
- Example: Suppose $Y = \mathbb{R}^m$ and $g \in C^1$. Then (BP) may be tackled using

$$\begin{array}{rl} \min\limits_{(x,y)\in X\times Y} & f(x,y) \\ \text{s.t.} & \nabla_y g(x,y) = 0 \end{array}.$$

- The above a *smooth* nonlinear programming problem if $f \in C^1$ and $g \in C^2$.

Traditional approaches

- Replace the lower-level (LL) problem by its optimality conditions.
- Example: Suppose $Y = \mathbb{R}^m$ and $g \in C^1$. Then (BP) may be tackled using

$$\begin{array}{rl} \min\limits_{(x,y)\in X\times Y} & f(x,y) \\ \text{s.t.} & \nabla_y g(x,y) = 0 \end{array}.$$

- The above a *smooth* nonlinear programming problem if $f \in C^1$ and $g \in C^2$.
- This is an equivalent formulation if $g(x,\cdot)$ is convex $\forall x \in X$, ...but only a relaxation otherwise.

Traditional approaches

- Replace the lower-level (LL) problem by its optimality conditions.
- Example: Suppose $Y = \mathbb{R}^m$ and $g \in C^1$. Then (BP) may be tackled using

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x,y) \\ \text{s.t.} \quad & \nabla_y g(x,y) = 0 \end{aligned}.$$

- The above a *smooth* nonlinear programming problem if $f \in C^1$ and $g \in C^2$.
- This is an equivalent formulation if $g(x,\cdot)$ is convex $\forall x \in X$, ...but only a relaxation otherwise.
- Other approaches: Reformulation via the value function

Shortcomings of existing works

- Smoothness and (strong) convexity are strong assumptions!
- LL objective may be nonsmooth and/or nonconvex in y (or merely convex).
- Example: Let $g(x, y) = \frac{1}{2}\|Ay - b\|^2 + x\|y\|_p^p$ with $p \in (0, 1]$
  - nonsmooth for any $p \in (0, 1]$
  - nonconvex when $p \in (0, 1)$
  - merely convex when $p = 1$

Shortcomings of existing works

- Smoothness and (strong) convexity are strong assumptions!
- LL objective may be nonsmooth and/or nonconvex in y (or merely convex).
- Example: Let $g(x, y) = \frac{1}{2}\|Ay - b\|^2 + x\|y\|_p^p$ with $p \in (0, 1]$
  - nonsmooth for any $p \in (0, 1]$
  - nonconvex when $p \in (0, 1)$
  - merely convex when $p = 1$
- **Research on nonsmooth nonconvex BP objective is scarce**.
  - How to design solution methods with theoretical guarantees?
  - How to define stationarity for general bilevel problems?

Outline

Value function approach

- We consider the value function defined as

$$v(x) = \min_{y \in Y} \; g(x, y).$$

- With this, (BP) is **equivalent** to

$$\begin{aligned} \min_{(x,y) \in X \times Y} \quad & f(x, y) \\ \text{s.t.} \quad & g(x, y) - v(x) \leq 0, \end{aligned} \tag{VFP}$$

Value function approach

- We consider the value function defined as

$$v(x) = \min_{y \in Y} \; g(x, y).$$

- With this, (BP) is **equivalent** to

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) - v(x) \le 0, \end{aligned} \tag{VFP}$$

- Advantage: No structural assumptions on g!

Value function approach

- We consider the value function defined as

$$v(x) = \min_{y \in Y} \; g(x, y).$$

- With this, (BP) is **equivalent** to

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x, y) \\ \text{s.t.} \quad & g(x, y) - v(x) \leq 0, \end{aligned} \tag{VFP}$$

- Advantage: No structural assumptions on g!
- Disadvantages:

  1. Absence of suitable constraint qualifications
     (When are local solutions of (VFP) stationary?)
  2. Lack of solution methods for (VFP) ← Focus of this work!

1. Constraint qualifications

Recall: If a local solution satisfies some constraint qualifications (CQ), then it is a stationary point.

- Usual CQs, such as the Mangasarian-Fromovitz CQ, are violated by feasible points of (VFP).

1. Constraint qualifications

Recall: If a local solution satisfies some constraint qualifications (CQ), then it is a stationary point.

- Usual CQs, such as the Mangasarian-Fromovitz CQ, are violated by feasible points of (VFP).
- Alternative: Consider the approximate bilevel program

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) - v(x) \leq \epsilon. \end{aligned} \tag{VFP$_\epsilon$}$$

  where $\epsilon > 0$.
- MFCQ is automatically satisfied by feasible points of (VFP$_\epsilon$).
- Local/global solutions of (VFP$_\epsilon$) are arbitrarily close to solution set of (VFP) (Lin et al., 2014, Ye et al., 2023).

2. Solution methods

- v is nonsmooth in general, even if g is smooth.
- Example: Let $g(x, y) = xy$ and $Y = [-1, 1]$. Then

$$v(x) = \min_{y \in Y} g(x, y) = -|x|.$$

- Hence, unfortunately, even if f and g are smooth, (VFP_ϵ) may be a **nonsmooth** nonlinear programming problem (NLP).

Outline

Proposed approach

- Target problem: The value function reformulation of (BP) with **nonsmooth** but Lipschitz continuous f and g:

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) - v(x) \le \epsilon, \end{aligned} \tag{VFP$_\epsilon$}$$

where $v(x) := \min_{y\in Y} g(x,y)$.

Proposed approach

- Target problem: The value function reformulation of (BP) with **nonsmooth** but Lipschitz continuous f and g:

$$\min_{(x,y)\in X\times Y} \quad f(x,y) \quad \text{s.t.} \quad g(x,y)-v(x)\le \epsilon, \tag{VFP$_\epsilon$}$$

  where $v(x) := \min_{y\in Y} g(x,y)$.
- Strategy:
  - Replace f and g with their smooth approximations.
  - Derive smooth approximations of the value function v.

Formal definition of smoothing functions

- Let $O \subseteq \mathbb{R}^d$ be an open set and let $\phi : O \to \mathbb{R}$ be a Lipschitz continuous function.

Definition (X. Chen, R. Womersley, and J. Ye, 2011)

We say that $\{\phi_\mu : \mu > 0\}$ is a family of smooth approximations for $\phi : O \to \mathbb{R}$ if $\phi_\mu : O \to \mathbb{R}$ is continuously differentiable and if

$$\lim_{z \to \bar{z}, \mu \to 0} \phi_\mu(z) = \phi(\bar{z}) \quad \forall z \in O.$$

A smooth approximation model of the bilevel problem

- We consider

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \qquad (\mathrm{VFP}_\epsilon^\mu)$$

  where f_μ, g_μ and v_μ are smooth approximations of f, g and v, resp.

- Smooth approximations f_μ and g_μ may be readily available.

A smooth approximation model of the bilevel problem

- We consider

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \qquad (\mathrm{VFP}_\epsilon^\mu)$$

  where f_μ, g_μ and v_μ are smooth approximations of f, g and v, resp.

- Smooth approximations f_μ and g_μ may be readily available.
- Problems
  1. How do we obtain approximations of v?
  2. How are the stationary points of $(\mathrm{VFP}_\epsilon^\mu)$ related to stationary points of (VFP_ϵ) as $\mu \to 0$?

A smooth approximation model of the bilevel problem

- We consider

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \qquad (\text{VFP}_\epsilon^\mu)$$

  where f_μ, g_μ and v_μ are smooth approximations of f, g and v, resp.

- Smooth approximations f_μ and g_μ may be readily available.
- Problems
  1. How do we obtain approximations of v?
  2. How are the stationary points of $(\text{VFP}_\epsilon^\mu)$ related to stationary points of (VFP_ϵ) as $\mu \to 0$?
- Goal: Derive smooth approximations[1] of the value function so that accumulation points of a sequence of stationary points $\{x_\mu : \mu > 0\}$ are stationary points of (VFP_ϵ).

[1] not smoothing algorithms!

Stationary points

Definition

Let ϕ be a Lipschitz continuous function on an open set $O \subseteq \mathbb{R}^n$.

1. The Clarke generalized directional derivative of ϕ at $\bar{x} \in O$ in the direction d, denoted by $\phi^\circ(x; d)$, is defined as

$$\phi^\circ(\bar{x}; d) = \limsup_{x \to \bar{x}, t \searrow 0} \frac{\phi(x + td) - \phi(x)}{t},$$

2. The Clarke generalized gradient of ϕ at $\bar{x}$, denoted by $\partial\phi(\bar{x})$, is given by

$$\partial\phi(\bar{x}) := \{\xi \in \mathbb{R}^n : \phi^\circ(\bar{x}; d) \geq \langle \xi, d \rangle \ \forall d \in \mathbb{R}^n\},$$

3. $\bar{x} \in X$ is a stationary point of

$$\min_{x \in X} \phi(x)$$

if $0 \in \partial\phi(\bar{x}) + N_X(\bar{x})$.

Stationary point of the value function reformulation

Recall our target problem:

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) - v(x) \le \epsilon. \end{aligned} \qquad (\text{VFP}_\epsilon)$$

Definition (Lin et al., 2014)

Let $(\bar{x},\bar{y})$ be a feasible point of (VFP_ϵ) with $\epsilon \ge 0$. We say that $(\bar{x},\bar{y})$ is a stationary point of (VFP_ϵ) if there exists $\lambda \ge 0$ such that

$$\begin{cases} 0 \in \partial f(\bar{x},\bar{y}) + \lambda\partial g(\bar{x},\bar{y}) - \lambda\partial v(\bar{x}) \times \{0\} + N_{X\times Y}(\bar{x},\bar{y}) \\ \lambda(g(\bar{x},\bar{y}) - v(\bar{x}) - \epsilon) = 0 \end{cases}$$

Goal

For the smoothly approximated problem,

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \qquad (\mathrm{VFP}_\epsilon^\mu)$$

1. How do we derive smooth approximations of v?
2. How do we ensure that if

$$\{x_\mu : \mu > 0\} \text{ is a sequence of stationary points of } (\mathrm{VFP}_\epsilon^\mu),$$

then accumulation points are stationary to the original problem (VFP_ϵ)?

An important requirement for smoothing approaches

- Consider the problem

$$\min_{x \in X} \phi(x)$$

  with nonsmooth ϕ.

- Smoothly approximate the problem as

$$\min_{x \in X} \phi_\mu(x).$$

- Let $\{\mu_k\}$ be a sequence such that $\mu_k \to 0$. Assume that
  - For each k, we can get a stationary point x^k, that is, $0 \in \nabla \phi_{\mu_k}(x^k) + N_X(x^k)$ as $k \to \infty$.
  - For simplicity, $x^k \to \bar{x}$.

An important requirement for smoothing approaches

- Consider the problem

$$\min_{x \in X} \phi(x)$$

with nonsmooth ϕ.

- Smoothly approximate the problem as

$$\min_{x \in X} \phi_{\mu}(x).$$

- Let $\{\mu_k\}$ be a sequence such that $\mu_k \to 0$. Assume that
  - For each k, we can get a stationary point x^k, that is, $0 \in \nabla\phi_{\mu_k}(x^k) + N_X(x^k)$ as $k \to \infty$.
  - For simplicity, $x^k \to \bar{x}$.
- Desired property: $0 \in \partial\phi(\bar{x}) + N_X(\bar{x})$.

Gradient consistency

Definition (X. Chen, R. Womersley, and J. Ye, 2011)

The family of smooth approximations $\{\phi_\mu : \mu > 0\}$ for ϕ satisfies the gradient consistent property at $\bar{z} \in O$ if

$$\emptyset \neq \limsup_{z \to \bar{z}, \mu \searrow 0} \nabla \phi_\mu(z) \subseteq \partial \phi(\bar{z}),$$

where

$$\limsup_{z \to \bar{z}, \mu \searrow 0} \nabla \phi_\mu(z) := \left\{ \xi \in \mathbb{R}^d : \exists \{z^k\}, \exists \{\mu_k\} \text{ s.t. } z^k \to \bar{z}, \mu_k \searrow 0 \text{ and } \nabla \phi_{\mu_k}(z^k) \to \xi \right\}$$

Significance of gradient consistency for bilevel problems

Proposition (A. and Takeda, 2024)

Let $\{\mu_k\}$ be a sequence of positive numbers with $\mu_k \searrow 0$. Suppose that

- $\{f_\mu : \mu > 0\}$, $\{g_\mu : \mu > 0\}$ and $\{v_\mu : \mu > 0\}$ satisfy the gradient consistent property;
- (x^k, y^k) is a stationary point of $(\text{VFP}^\mu_\epsilon)$ with $\mu = \mu_k$;
- Let λ_k denote the corresponding Lagrange multiplier.

If $\{(x^k, y^k, \lambda_k)\}$ is bounded, then its accumulation points are stationary points of (VFP_ϵ).

Summary

- We consider the model

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \tag{$\mathrm{VFP}_\epsilon^\mu$}$$

- **Assumption:** Smoothing functions f_μ and g_μ that possess gradient consistent property are available.

Goals

- Derive a smooth approximation v_μ of $v = \min_{y\in Y} g(\cdot, y)$.

- Establish gradient consistency:

$$\emptyset \neq \limsup_{x\to\bar{x},\mu\searrow 0} \nabla v_\mu(x) \subseteq \partial v(\bar{x}),$$

Summary

- We consider the model

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \le \epsilon. \end{aligned} \qquad (\text{VFP}_\epsilon^\mu)$$

- **Assumption:** Smoothing functions f_μ and g_μ that possess gradient consistent property are available.

Goals

- Derive a smooth approximation v_μ of $v = \min_{y\in Y} g(\cdot, y)$.
- Characterize the elements of ∂v.
- Establish gradient consistency:

$$\emptyset \neq \limsup_{x\to\bar{x},\mu\searrow 0} \nabla v_\mu(x) \subseteq \partial v(\bar{x}),$$

Outline

Smoothing approach 1: Quadratic regularization

- Consider $\{\tilde{g}_\mu : \mu > 0\}$ where

$$\tilde{g}_\mu(x, y) := g_\mu(x, y) + \frac{\mu}{2}\|y\|^2,$$

and define

$$v_\mu(x) := \min_{y \in Y} \tilde{g}_\mu(x, y) \quad \text{and} \quad S_\mu(x) := \operatorname*{arg\,min}_{y \in Y} \tilde{g}_\mu(x, y).$$

Smoothing approach 1: Quadratic regularization

- Consider $\{\tilde{g}_\mu : \mu > 0\}$ where

$$\tilde{g}_\mu(x, y) := g_\mu(x, y) + \frac{\mu}{2}\|y\|^2,$$

and define

$$v_\mu(x) := \min_{y \in Y} \tilde{g}_\mu(x, y) \quad \text{and} \quad S_\mu(x) := \operatorname*{arg\,min}_{y \in Y} \tilde{g}_\mu(x, y).$$

- Questions: When is v_μ smooth? When do we have

$$\lim_{x \to \bar{x}, \mu \searrow 0} v_\mu(x) = v(\bar{x}) \ ?$$

Smoothing approach 1: Quadratic regularization

- Consider $\{\tilde{g}_\mu : \mu > 0\}$ where

$$\tilde{g}_\mu(x,y) := g_\mu(x,y) + \frac{\mu}{2}\|y\|^2,$$

and define

$$v_\mu(x) := \min_{y\in Y} \tilde{g}_\mu(x,y) \quad \text{and} \quad S_\mu(x) := \operatorname*{arg\,min}_{y\in Y} \tilde{g}_\mu(x,y).$$

- Questions: When is v_μ smooth? When do we have

$$\lim_{x\to\bar{x},\mu\searrow 0} v_\mu(x) = v(\bar{x}) \text{ ?}$$

Under what conditions can we achieve

$$\lim_{x\to\bar{x},\mu\searrow 0}\left(\min_{y\in Y} \tilde{g}_\mu(x,y)\right) \overset{?}{=} \min_{y\in Y} \underbrace{\left(\lim_{x\to\bar{x},\mu\searrow 0} \tilde{g}_\mu(x,y)\right)}_{=g(\bar{x},y)} \text{ ?}$$

Smoothing approach 1: Quadratic regularization

- Consider $\{\tilde{g}_\mu : \mu > 0\}$ where

$$\tilde{g}_\mu(x,y) \coloneqq g_\mu(x,y) + \frac{\mu}{2}\|y\|^2,$$

and define

$$v_\mu(x) \coloneqq \min_{y\in Y} \tilde{g}_\mu(x,y) \quad \text{and} \quad S_\mu(x) \coloneqq \arg\min_{y\in Y} \tilde{g}_\mu(x,y).$$

- Questions: When is v_μ smooth? When do we have

$$\lim_{x\to\bar{x},\mu\searrow 0} v_\mu(x) = v(\bar{x}) \ ?$$

Under what conditions can we achieve

$$\lim_{x\to\bar{x},\mu\searrow 0} v_\mu(x) \overset{\text{def}}{=} \lim_{x\to\bar{x},\mu\searrow 0} \left(\min_{y\in Y} \tilde{g}_\mu(x,y)\right) \overset{?}{=} \min_{y\in Y} \underbrace{\left(\lim_{x\to\bar{x},\mu\searrow 0} \tilde{g}_\mu(x,y)\right)}_{=g(\bar{x},y)} \ ?$$

An important tool

Definition

A function $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with values $h(y, z)$ is level-bounded in y locally uniform in z if for any $z' \in \mathbb{R}^n$ and $M \in \mathbb{R}$, there exists an open ball B around z' such that

$$\bigcup_{z \in B} \{y \in \mathbb{R}^m : h(y, z) \leq M\}$$

is bounded. We also say that h is uniformly level bounded.

An important tool

Theorem (Rockafellar and Wets 1998)

Let $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function that is level-bounded in $y \in \mathbb{R}^m$ locally uniform in $z \in \mathbb{R}^d$. Define

$$v(z) := \min_{y \in \mathbb{R}^m} h(y, z) \quad \text{and} \quad S(z) := \arg\min_{y \in \mathbb{R}^m} h(y, z)$$

and let $\bar{z} \in \mathbb{R}^d$.

(a) If there exists $\bar{y} \in S(\bar{z})$ such that $h(\bar{y}, \cdot)$ is continuous on a set U containing $\bar{z}$, then v is continuous on U; and

(b) If v is continuous on a set U containing $\bar{z}$, $\{z^k\} \subseteq U$ such that $z^k \to \bar{z}$ and $\{y^k\}$ is a sequence such that $y^k \in S(z^k)$ for all k, then $\{y^k\}$ is bounded and its accumulation points lie on $S(\bar{z})$.

An important tool

Theorem (Rockafellar and Wets 1998)

Let $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function that is level-bounded in $y \in \mathbb{R}^m$ locally uniform in $z \in \mathbb{R}^d$. Define

$$v(z) := \min_{y \in \mathbb{R}^m} h(y, z) \quad \text{and} \quad S(z) := \operatorname*{arg\,min}_{y \in \mathbb{R}^m} h(y, z)$$

and let $\bar{z} \in \mathbb{R}^d$.

(a) If there exists $\bar{y} \in S(\bar{z})$ such that $h(\bar{y}, \cdot)$ is continuous on a set U containing $\bar{z}$, then v is continuous on U; and

(b) If v is continuous on a set U containing $\bar{z}$, $\{z^k\} \subseteq U$ such that $z^k \to \bar{z}$ and $\{y^k\}$ is a sequence such that $y^k \in S(z^k)$ for all k, then $\{y^k\}$ is bounded and its accumulation points lie on $S(\bar{z})$.

Note: continuity of v at $\bar{z}$ means that

$$\min_{y \in \mathbb{R}^m} h(y, \bar{z}) \stackrel{\text{def}}{=} \lim_{z \to \bar{z}} v(z)$$

An important tool

Theorem (Rockafellar and Wets 1998)

Let $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function that is level-bounded in $y \in \mathbb{R}^m$ locally uniform in $z \in \mathbb{R}^d$. Define

$$v(z) := \min_{y \in \mathbb{R}^m} h(y,z) \quad \text{and} \quad S(z) := \operatorname*{arg\,min}_{y \in \mathbb{R}^m} h(y,z)$$

and let $\bar{z} \in \mathbb{R}^d$.

(a) If there exists $\bar{y} \in S(\bar{z})$ such that $h(\bar{y}, \cdot)$ is continuous on a set U containing $\bar{z}$, then v is continuous on U; and

(b) If v is continuous on a set U containing $\bar{z}$, $\{z^k\} \subseteq U$ such that $z^k \to \bar{z}$ and $\{y^k\}$ is a sequence such that $y^k \in S(z^k)$ for all k, then $\{y^k\}$ is bounded and its accumulation points lie on $S(\bar{z})$.

Note: continuity of v at $\bar{z}$ means that

$$\min_{y \in \mathbb{R}^m} h(y, \bar{z}) \overset{\text{def}}{=} \lim_{z \to \bar{z}} v(z) \overset{\text{def}}{=} \lim_{z \to \bar{z}} \left(\min_{y \in \mathbb{R}^m} h(y,z) \right)$$

An important tool

Theorem (Rockafellar and Wets 1998)

Let $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function that is level-bounded in $y \in \mathbb{R}^m$ locally uniform in $z \in \mathbb{R}^d$. Define

$$v(z) := \min_{y \in \mathbb{R}^m} h(y, z) \quad \text{and} \quad S(z) := \operatorname*{arg\,min}_{y \in \mathbb{R}^m} h(y, z)$$

and let $\bar{z} \in \mathbb{R}^d$.

(a) If there exists $\bar{y} \in S(\bar{z})$ such that $h(\bar{y}, \cdot)$ is continuous on a set U containing $\bar{z}$, then v is continuous on U; and

(b) If v is continuous on a set U containing $\bar{z}$, $\{z^k\} \subseteq U$ such that $z^k \to \bar{z}$ and $\{y^k\}$ is a sequence such that $y^k \in S(z^k)$ for all k, then $\{y^k\}$ is bounded and its accumulation points lie on $S(\bar{z})$.

Note: If h is continuous, continuity of v at $\bar{z}$ means that

$$\min_{y \in \mathbb{R}^m} h(y, \bar{z}) \stackrel{\text{def}}{=} \lim_{z \to \bar{z}} v(z) \stackrel{\text{def}}{=} \lim_{z \to \bar{z}} \left(\min_{y \in \mathbb{R}^m} h(y, z) \right)$$

An important tool

Theorem (Rockafellar and Wets 1998)

Let $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function that is level-bounded in $y \in \mathbb{R}^m$ locally uniform in $z \in \mathbb{R}^d$. Define

$$v(z) := \min_{y \in \mathbb{R}^m} h(y, z) \quad \text{and} \quad S(z) := \operatorname*{arg\,min}_{y \in \mathbb{R}^m} h(y, z)$$

and let $\bar{z} \in \mathbb{R}^d$.

(a) If there exists $\bar{y} \in S(\bar{z})$ such that $h(\bar{y}, \cdot)$ is continuous on a set U containing $\bar{z}$, then v is continuous on U; and

(b) If v is continuous on a set U containing $\bar{z}$, $\{z^k\} \subseteq U$ such that $z^k \to \bar{z}$ and $\{y^k\}$ is a sequence such that $y^k \in S(z^k)$ for all k, then $\{y^k\}$ is bounded and its accumulation points lie on $S(\bar{z})$.

Note: If h is continuous, continuity of v at $\bar{z}$ means that

$$\min_{y \in \mathbb{R}^n} \left(\lim_{z \to \bar{z}} h(y, z) \right) = \min_{y \in \mathbb{R}^m} h(y, \bar{z}) \stackrel{\text{def}}{=} \lim_{z \to \bar{z}} v(z) \stackrel{\text{def}}{=} \lim_{z \to \bar{z}} \left(\min_{y \in \mathbb{R}^m} h(y, z) \right)$$

Smoothing approach 1: Quadratic regularization

Assumption A

Let O_X be an open set containing X and $\Omega \subseteq O_X$.

1. $g_\mu(x, \cdot)$ is **convex** $\forall x \in O_X$ and g_μ is uniformly bounded on $\mathrm{cl}(\Omega) \times Y$.
2. $g_\mu(x, y)$ is **continuous** as a function of $(x, y, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++}$.
3. $g_\mu \to g$ **uniformly** on $\mathrm{cl}(\Omega) \times Y$ as $\mu \to 0$ for any $\Omega \subseteq O_X$.
4. g is **level-bounded in** y **locally uniform in** x.

Theorem 1 (A. and Takeda, 2024)

Under Assumption A1-A4, $\{v_\mu : \mu > 0\}$ is a family of smooth approximations for v and $\nabla v_\mu(x) = \nabla_x g_\mu(x, S_\mu(x))$.

Some comments

- If $g(x,\cdot)$ is convex, then its Moreau envelope, i.e.,

$$M_{g(x,\cdot)}(y) = \min_{z \in Y} \; g(x,z) + \frac{1}{2\mu}\|z - y\|^2$$

is a convex differentiable function that **converges uniformly** to $g(x,y)$.

We can set $g_\mu(x,y) = M_{g(x,\cdot)}(y)$ and Assumptions A1, A3 are satisfied.

Some comments

- If $g(x,\cdot)$ is convex, then its Moreau envelope, i.e.,

$$M_{g(x,\cdot)}(y) = \min_{z \in Y} \; g(x,z) + \frac{1}{2\mu}\|z-y\|^2$$

  is a convex differentiable function that **converges uniformly** to $g(x,y)$.

  We can set $g_\mu(x,y) = M_{g(x,\cdot)}(y)$ and Assumptions A1, A3 are satisfied.

- In hyperparameter learning, we consider the LL objective

$$g(x,y) = \ell(y) + \sum_{i=1}^{n} x_i p_i(y),$$

  with constraint sets $Y = \mathbb{R}^m$ and $X = [\varepsilon_1,\infty) \times \cdots \times [\varepsilon_n,\infty)$, where $\varepsilon_i > 0$ for all i.

  If one of the p_i's is coercive, then Assumption A4 is satisfied.

Smoothing approach 2: Entropic regularization – Compact case

[1]Inspired by Lin et al. (2014), Fang and Wu (1996), and Li and Fang (1997).

Smoothing approach 2: Entropic regularization – Compact case

- Recall the approximation

$$\max\{y_1, y_2, \ldots, y_r\} \approx \mu \ln \sum_{i=1}^{r} \exp(\mu^{-1} y_i).$$

[1] Inspired by Lin et al. (2014), Fang and Wu (1996), and Li and Fang (1997).

Smoothing approach 2: Entropic regularization — Compact case

- Recall the approximation

$$\max\{y_1, y_2, \ldots, y_r\} \approx \mu \ln \sum_{i=1}^{r} \exp(\mu^{-1} y_i).$$

- We propose the following approximation[1] of v:

$$v_\mu(x) := -\mu \ln \left(\int_Y \exp\left(-\mu^{-1} g_\mu(x, y)\right) dy \right)$$

[1] Inspired by Lin et al. (2014), Fang and Wu (1996), and Li and Fang (1997).

Smoothing approach 2: Entropic regularization – Compact case

Assumption A

Let O_X be an open set containing X and $\Omega \subseteq O_X$.

1. $g_\mu(x, \cdot)$ is **convex** $\forall x \in O_X$ and g_μ is uniformly bounded on $\text{cl}(\Omega) \times Y$.
2. $g_\mu(x, y)$ is **continuous** as a function of $(x, y, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++}$.
3. $g_\mu \to g$ **uniformly** on $\text{cl}(\Omega) \times Y$ as $\mu \to 0$ for any $\Omega \subseteq O_X$.
4. g is **level-bounded in** y **locally uniform in** x.

Smoothing approach 2: Entropic regularization – Compact case

- Recall the approximation

$$\max\{y_1, y_2, \ldots, y_r\} \approx \mu \ln \sum_{i=1}^{r} \exp(\mu^{-1} y_i).$$

- We propose the following approximation[1] of v:

$$v_\mu(x) := -\mu \ln \left(\int_Y \exp\left(-\mu^{-1} g_\mu(x, y)\right) dy \right)$$

Theorem 2 (A. and Takeda, 2024)

Under Assumption A3, $\{v_\mu : \mu > 0\}$ is a family of smooth approximations for v provided that Y is compact.

[1] Inspired by Lin et al. (2014), Fang and Wu (1996), and Li and Fang (1997).

Proof ingredients:

1. Leibniz rule
2. Mean-value theorem for Lipschitz continuous functions
3. Integral estimates:

   For any $\tau \in (0,1)$, there exists $\delta > 0$ such that for any $\mu \in (0,\delta)$ and $x \in \mathrm{cl}(\Omega)$,

$$\begin{aligned} \tau(\mu\,\mathrm{vol}(Y))^{\mu} \max_{y\in Y} \exp(-g(x,y)) &\leq \left(\int_Y \exp\left(-\mu^{-1} g(x,y)\right) dy\right)^{\mu} \\ &\leq \mathrm{vol}(Y)^{\mu} \max_{y\in Y} \exp(-g(x,y)) \end{aligned}$$

Smoothing approach 2: Entropic regularization – Unbounded case

- When Y is an unbounded closed set, we consider

$$v_\mu(x) := -\mu \ln \left(\int_{Y_\mu} \exp\left(-\mu^{-1} g_\mu(x,y)\right) dy \right), \tag{1}$$

where Y_μ is a compact set such that $Y_\mu \nearrow Y$ as $\mu \searrow 0$.

Theorem 3 (A. and Takeda, 2024)

Under Assumption A3-A4 and assuming that $\mu \ln \mathrm{vol}(Y_\mu) \to 0$ as $\mu \searrow 0$, $\{v_\mu : \mu > 0\}$ is a family of smooth approximations for v.

Goals

- ~~Derive a smooth approximation v_μ of v~~.
- Characterize the elements of ∂v.
- Establish gradient consistency:

$$\emptyset \neq \limsup_{x \to \bar{x}, \mu \searrow 0} \nabla v_\mu(x) \subseteq \partial v(\bar{x}),$$

Outline

Smooth case, compact constraint set

Danskin's Theorem (Danskin, 1967)

Let $g : \mathbb{R}^n \times Y \to \mathbb{R}$ be given, where Y is a compact subset of $\mathbb{R}^m$. Suppose that for a neighborhood Ω of $\bar{x} \in \mathbb{R}^n$, the derivative $\nabla_x g(x, y)$ exists and is continuous (jointly) as a function of $(x, y) \in \Omega \times Y$. Then

$$\partial v(\bar{x}) = \text{co}\{\nabla_x g(\bar{x}, \bar{y}) : \bar{y} \in S(\bar{x})\},$$

where $S : \mathbb{R}^n \rightrightarrows Y$ is given by

$$S(x) := \operatorname*{arg\,min}_{y \in Y} g(x, y).$$

Nonsmooth case, compact constraint set

Danskin-type theorem for nonsmooth functions (Bertsekas, 1971)

Let $g : \mathbb{R}^n \times Y \to \mathbb{R}$ be given, where Y is a compact subset of $\mathbb{R}^m$. Suppose that

- for a neighborhood Ω of $\bar{x} \in \mathbb{R}^n$, g is continuous on $\Omega \times Y$; and
- for every $y \in Y$, the function $g(\cdot, y)$ is a concave function on $\mathbb{R}^n$

Then v is concave on $\mathbb{R}^n$ and

$$\partial v(\bar{x}) = P(\bar{x}) := \text{co}\{\xi \in \mathbb{R}^n : \xi \in \partial_x g(\bar{x}, \bar{y}) \text{ and } \bar{y} \in S(\bar{x})\}.$$

Generalizations

Let $\bar{x} \in O_X$ and Ω a neighborhood of $\bar{x}$.

Theorem 4 (A. and Takeda, 2024)

Suppose that Assumption A4 holds, i.e., g is level-bounded in y locally uniform in x. Moreover, suppose that any one of the following conditions hold:

1. $\nabla_x g(x, y)$ exists and is continuous on $\Omega \times Y$;
2. $g(\cdot, y)$ is ρ-weakly concave on Ω for every $y \in O_Y$;
3. g is convex in (x, y) and $\partial g(\bar{x}, \bar{y}) = \partial_x g(\bar{x}, \bar{y}) \times \partial_y g(\bar{x}, \bar{y})$ for every $\bar{y} \in S(\bar{x})$.

Then

$$\partial v(\bar{x}) = \mathrm{co}\{\xi \in \mathbb{R}^n : \xi \in \partial_x g(\bar{x}, \bar{y}) \text{ and } \bar{y} \in S(\bar{x})\}.$$

Goals

- ~~Derive a smooth approximation v_μ of v.~~
- ~~Characterize the elements of ∂v.~~
- Establish gradient consistency:

$$\emptyset \neq \limsup_{x \to \bar{x}, \mu \searrow 0} \nabla v_\mu(x) \subseteq \partial v(\bar{x}),$$

Outline

Gradient consistency for smoothing by quadratic regularization

Theorem 5 (A. and Takeda, 2024)

In addition to Assumption A, suppose that $\{g_\mu : \mu > 0\}$ satisfies the gradient consistent property. Then v satisfies the gradient consistent property at $\bar{x}$ if one of the following conditions holds:

1. $\nabla_x g(x, y)$ exists and is continuous on $\Omega \times Y$;
2. $g(\cdot, y)$ is ρ-weakly concave on Ω for every $y \in O_Y$;
3. g is convex in (x, y) and $\partial g(\bar{x}, \bar{y}) = \partial_x g(\bar{x}, \bar{y}) \times \partial_y g(\bar{x}, \bar{y})$ for every $\bar{y} \in S(\bar{x})$.

Gradient consistency for smoothing by quadratic regularization

Theorem 5 (A. and Takeda, 2024)

In addition to Assumption A, suppose that $\{g_\mu : \mu > 0\}$ satisfies the gradient consistent property. Then v satisfies the gradient consistent property at $\bar{x}$ if one of the following conditions holds:

1. $\nabla_x g(x, y)$ exists and is continuous on $\Omega \times Y$;
2. $g(\cdot, y)$ is ρ-weakly concave on Ω for every $y \in O_Y$;
3. g is convex in (x, y) and $\partial g(\bar{x}, \bar{y}) = \partial_x g(\bar{x}, \bar{y}) \times \partial_y g(\bar{x}, \bar{y})$ for every $\bar{y} \in S(\bar{x})$.

- Proof. Using uniform level-boundedness and gradient consistency, show that

$$\pi_x(\partial g(\bar{x}, \bar{y})) \subseteq \partial_x g(\bar{x}, \bar{y}) \quad \forall \bar{y} \in S(\bar{x})$$
$$\Longrightarrow \quad \emptyset \neq \limsup_{x \to \bar{x}, \mu \to 0} \nabla v_\mu(x) \subseteq \bigcup_{\bar{y} \in S(\bar{x})} \partial_x g(\bar{x}, \bar{y}).$$

Use Danskin's theorem.

Gradient consistency for smoothing by entropic regularization - Compact case

Theorem 6 (A. and Takeda, 2024)

In addition to Assumption A3, suppose that

(a) there exists a neighborhood Ω of $\bar{x} \in O_X$ such that $\partial_x g(\cdot,\cdot)$ is upper semicontinuous on $\Omega \times O_Y$; and

(b) $\mathrm{dist}(\nabla_x g_\mu(x,\cdot), \partial_x g(x,\cdot))$ converges to 0 uniformly on Y as $(x,\mu) \to (\bar{x}, 0)$

Then v satisfies the gradient consistent property at $\bar{x}$ if one of the following conditions holds:

1. $\nabla_x g(x,y)$ exists and is continuous on $\Omega \times Y$;
2. $g(\cdot,y)$ is ρ-weakly concave on Ω for every $y \in O_Y$;
3. g is convex in (x,y) and $\partial g(\bar{x},\bar{y}) = \partial_x g(\bar{x},\bar{y}) \times \partial_y g(\bar{x},\bar{y})$ for every $\bar{y} \in S(\bar{x})$.

[3] A similar result holds when Y is unbounded, but additional technical assumptions are needed.

Proof ingredients:

1. Uniform convergence + Bounded convergence theorem
2. Heine-Borel Theorem
3. Jensen's inequality
4. Rademacher's Theorem

Summary

- *Nonsmooth nonconvex* bilevel optimization is an area of optimization with lots of open problems!
- This work proposed two *theoretical frameworks* for deriving smooth approximations of the value function that possess gradient consistent property.
- Future works
  - Deriving other smooth approximations
  - Extensions to non-Lipschitz continuous functions
  - Specializing the results to min-max problems*
  - Numerical implementations of the smoothing approaches:*

$$\begin{aligned} \min_{(x,y)\in X\times Y} \quad & f_\mu(x,y) \\ \text{s.t.} \quad & g_\mu(x,y) - v_\mu(x) \leq \epsilon. \end{aligned} \qquad (\text{VFP}^\mu_\epsilon)$$

Thank you for listening!

Main reference: Alcantara, Jan Harold and Takeda, Akiko, “Theoretical smoothing frameworks for general nonsmooth bilevel problems”, arXiv:2401.17852 (2024).

References I

- Jan Harold Alcantara, Chieu Thanh Nguyen, Takayuki Okuno, Akiko Takeda, and Jein-Shan Chen. Unified smoothing approach for best hyperparameter selection problem using a bilevel optimization strategy, 2021. arXiv:2110.12630.
- Chunhui Chen and Olvi L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. Mathematical Programming, 71:51–70, 1995.
- Xiaojun Chen. Smoothing methods for nonsmooth, nonconvex minimization. Mathematical Programming Series B, 134:71–99, 2012.
- Xiaojun Chen, Robert S. Womersley, and Jane J. Ye. Minimizing the condition number of a gram matrix. SIAM Journal on Optimization, 21:127–148, 2011.
- Frank H. Clarke. Optimization and Nonsmooth Analysis. Wiley-Interscience, New York, 1983.
- Shu-Cherng Fang and Soon-Yi Wu. Solving min-max problems and linear semi-infinite programs. Computers and Mathematics with Applications, 32(6):87–93, 1996.
- Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming, 2018. arXiv:1802.02246.

References II

- Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 4882–4892. PMLR, 18–24 Jul 2021.
- Xing-Si Li and Shu-Cherng Fang. On the entropic regularization method for solving min-max problems with applications. Mathematical Methods of Operations Research, 46:119–130, 1997.
- Gui-Hua Lin, Mengwei Xu, and Jane J. Ye. On solving simple bilevel programs with a nonconvex lower level program. Mathematical Programming Series A, 144:277–305, 2014.
- R. Tyrrell Rockafellar and Roger J-B Wets. Variational Analysis, volume 317 of Grundlehren der MathematischenWissenschaften. Springer, Berlin, 1998.
- Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method, 2022. arXiv:2302.05185v4.
- Jane J. Ye and Zhu Daoli. Optimality conditions for bilevel programming problems. Optimization, 33:9–27, 1995.
- Jane J. Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. Mathematical Programming, 198: 1583–1616, 2023.

Nonsmooth MFCQ

Consider the problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g(x) \leq 0 \\ & x \in X. \end{array} \tag{2}$$

Definition (Lin, Xu & Ye, 2014)

Let $\bar{x}$ be a feasible point of (2). We say that the nonsmooth MFCQ holds at $\bar{x}$ if either $g(\bar{x}) < 0$ or $g(\bar{x}) = 0$ but there exists $d \in \operatorname{int} T_X(\bar{x})$ such that

$$v^\top d < 0 \quad \forall v \in \partial g(\bar{x}).$$

- If $\operatorname{int} T_X(\bar{x}) \neq \emptyset$, the latter condition is equivalent to having

$$0 \notin \partial g(\bar{x}) + N_X(\bar{x}).$$

General smoothing techniques

Suppose that h is a nonsmooth Lipschitz continuous function.

1. **Mollifiers**
   Given a mollifier ϕ, i.e., a compactly supported function with $\int_{\mathbb{R}^n} \phi(z)dz = 1$, we define

$$h_\mu(x) = \int_{\mathbb{R}^n} h(x - z)\phi_\mu(z)dz$$

   where $\phi_\mu(z) = \frac{1}{\mu^n}\phi\left(\frac{z}{\mu}\right)$.

   - h_μ is a smooth approximation of g:

$$\lim_{x \to \bar{x}, \mu \searrow 0} h_\mu(x) = h(\bar{x})$$

   - satisfies gradient consistency:

$$\emptyset \neq \limsup_{z \to \bar{z}, \mu \searrow 0} \nabla h_\mu(z) \subseteq \partial h(\bar{z}). \tag{3}$$

- Advantage: No restrictive assumptions.
- Disadvantage: For the value function, smoothing via mollifiers is too complex:

$$v_\mu(x) = \int_{\mathbb{R}^n} v(x-z)\phi_\mu(z)dz$$
$$= \int_{\mathbb{R}^n} \min_{y \in Y} g(x-z, y)\phi_\mu(z)dz.$$

2 **Infimal convolution**

Suppose that h is convex and ϕ is an L-smooth convex function. Define $\phi_\mu(\cdot) := \mu\phi\left(\frac{\cdot}{\mu}\right)$, and

$$h_\mu(x) := \min_{z\in\mathbb{R}^n} h(z) + \phi_\mu(x-z) = (h \Box \phi_\mu)(x)$$

- h_μ is L/μ-smooth with

$$\nabla h_\mu(x) = \nabla\phi\left(\frac{x - p_\mu(x)}{\mu}\right)$$

where

$$p_\mu(x) := \operatorname*{arg\,min}_{z\in\mathbb{R}^n} h(z) + \phi_\mu(x-z) = (h \Box \phi_\mu)(x)$$

- Gradient consistency holds.

- Disadvantage: Has additional requirement on h.
  - i For the value function v to be convex, one sufficient condition is for $g(x,y)$ to be jointly convex on (x,y);
  - ii $-v$ is convex if $g(\cdot,y)$ is concave for each $y \in Y$.
- Advantage: May be easier to compute for the value function
  - i $v_\mu(x) = \min_{z\in\mathbb{R}^n} v(z) + \mu\phi\left(\frac{x-z}{\mu}\right) = \min_{z\in\mathbb{R}^n}\min_{y\in Y} g(x,y) + \mu\phi\left(\frac{x-z}{\mu}\right)$
  - ii $v_\mu(x) := -(-v\Box\phi_\mu)(x) = \max_{z\in\mathbb{R}^n}\min_{y\in Y} g(z,y) - \mu\phi\left(\frac{x-z}{\mu}\right)$

Summary

- Smoothing via mollifiers is generally applicable but may be computationally intractable
  - Involves minimization and integration
- Smoothing via infimal convolution is applicable for special convex/concave cases
  - Involves double optimization
- Both satisfy gradient consistency.