Functional Bilevel Optimization: Theory and Algorithms

RKHS Seminars (METU)

Michael Arbel

INRIA, Grenoble Rhône-Alpes, France

February 13, 2025

Outline

Motivation: Objectives and challenges in bilevel optimization

Part I: Functional bilevel optimization

Part II: Towards a learning theory for Kernel Bilevel optimization

Outline

Motivation: Objectives and challenges in bilevel optimization

Part I: Functional bilevel optimization

Part II: Towards a learning theory for Kernel Bilevel optimization

Bilevel Optimization (BO) in machine learning

$$\min_{\omega \in \Omega} \mathscr{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$ ← Upper-level (e.g., Validation loss)

$$\text{s.t.} \quad \theta^\star_\omega \in \arg\min_{\theta \in \Theta} L_{in}(\omega, \theta)$$ ← Lower-level (e.g., Training loss)

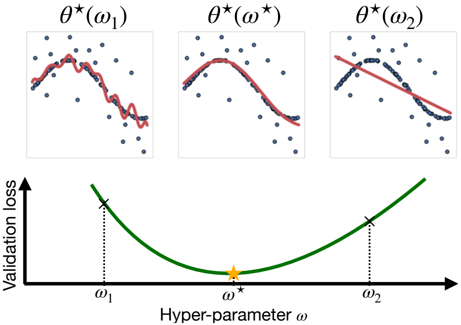
Goal: Minimizing $\mathscr{L}(\omega)$ using (approximate) gradient methods.

Bilevel Optimization (BO) in machine learning

$$\min_{\omega \in \Omega} \mathscr{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$ Upper-level (e.g., Validation loss)

$$\text{s.t.} \quad \theta^\star_\omega \in \arg\min_{\theta \in \Theta} L_{in}(\omega, \theta)$$ Lower-level (e.g., Training loss)

Goal: Minimizing $\mathscr{L}(\omega)$ using (approximate) gradient methods.



Bilevel Optimization (BO) in machine learning

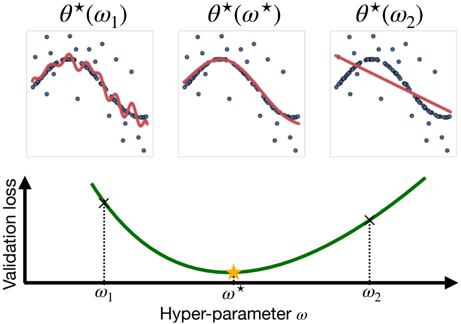
$$\min_{\omega \in \Omega} \mathscr{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$
Upper-level (e.g., Validation loss)

$$\text{s.t.} \quad \theta^\star_\omega \in \arg\min_{\theta \in \Theta} L_{in}(\omega, \theta)$$
Lower-level (e.g., Training loss)

Goal: Minimizing $\mathscr{L}(\omega)$ using (approximate) gradient methods.

Many machine learning applications:

- Hyper-parameter Optimization [Franceschi et al., 2018]
- Meta-learning [Rajeswaran et al., 2019]
- Model-based Reinforcement learning [Nikishin et al., 2022]
- GANs [Zhang et al., 2022]
- Dictionary learning [Mairal et al., 2011]



General Bilevel problems are (very) hard

Harder than NP-Hard

General bilevel problems are provably harder than general optimization [Bolte et al., 2024].

Any Hope?

Specialized methods can find solutions efficiently when:

- The lower-level problem is **strongly convex**,
- and has **finite-dimensional variables**.

To tame the complexity of bilevel problems, we must exploit every structural advantage!

Strongly-convex lower-level with finite dimensional lower variables

A **manageable**, but restrictive, setting

$$\min_{\omega \in \mathbb{R}^d} \mathcal{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$
$$\text{s.t.} \quad \theta^\star_\omega = \arg\min_{\theta \in \mathbb{R}^p} L_{in}(\omega, \theta)$$

$\Big\downarrow$ Chain rule

$$\nabla \mathcal{L}(\omega) = \partial_1 L_{out}(\omega, \theta^\star_\omega) + \nabla \theta^\star_\omega \partial_2 L_{out}(\omega, \theta^\star_\omega)$$

$\Big\downarrow$ IFT

$$\nabla \theta^\star_\omega \overbrace{\partial^2_{2,2} L_{in}(\omega, \theta^\star_\omega)}^{p \times p \text{ matrix}} = -\overbrace{\partial^2_{1,2} L_{in}(\omega, \theta^\star_\omega)}^{d \times p \text{ matrix}}$$

Key ingredient: Implicit differentiation

- Strong convexity guarantees existence and uniqueness of the inner-level $\theta^\star_\omega$.
- **Implicit function theorem (IFT):** $\nabla \theta^\star_\omega$ defined by a linear system.

Strongly-convex lower-level with finite dimensional lower variables

A **manageable**, but restrictive, setting

$$\min_{\omega \in \mathbb{R}^d} \mathcal{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$

$$\text{s.t.} \quad \theta^\star_\omega = \arg\min_{\theta \in \mathbb{R}^p} L_{in}(\omega, \theta)$$

$$\Big\downarrow \text{ Chain rule}$$

$$\nabla \mathcal{L}(\omega) = \partial_1 L_{out}(\omega, \theta^\star_\omega) + \nabla \theta^\star_\omega \partial_2 L_{out}(\omega, \theta^\star_\omega)$$

$$\Big\downarrow \text{ IFT}$$

$$\nabla \theta^\star_\omega \overbrace{\partial^2_{2,2} L_{in}(\omega, \theta^\star_\omega)}^{p \times p \text{ matrix}} = -\overbrace{\partial^2_{1,2} L_{in}(\omega, \theta^\star_\omega)}^{d \times p \text{ matrix}}$$

Key ingredient: Implicit differentiation

- Strong convexity guarantees existence and uniqueness of the inner-level $\theta^\star_\omega$.
- **Implicit function theorem (IFT):** $\nabla \theta^\star_\omega$ defined by a linear system.

Efficient algorithms + Theoretical guarantees

- Scalable algorithms (AID, ITD, ...) [Lorraine et al., 2020, Franceschi et al., 2017]
- Near optimal convergence guarantees in various settings (deterministic/stochastic): [Ghadimi and Wang, 2018, Ji et al., 2021, Arbel and Mairal, 2021, Dagréou et al., 2022].

Strongly-convex lower-level with finite dimensional lower variables

A manageable, but **restrictive**, setting

$$\min_{\omega \in \mathbb{R}^d} \mathscr{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$
$$\text{s.t.} \quad \theta^\star_\omega = \arg\min_{\theta \in \mathbb{R}^p} L_{in}(\omega, \theta)$$

Inner variable θ often represents the parameters of some predictive model f_θ, ex:

$$L_{in}(\omega, \theta) = \mathbb{E}\left[\|y - f_\theta(x)\|^2 + e^\omega \|f_\theta(x)\|^2\right]$$

- Strong convexity of $\theta \mapsto L_{in}(\omega, \theta)$ **restricts** to linear model, i.e.:

$$f_\theta(x) = \theta^\top \psi(x)$$

- Sophisticated models f_θ, (e.g. neural networks) $\implies$ $\theta \mapsto L_{in}(\omega, \theta)$ is **non-convex.**

Should we go non-convex?

Non ~~Strongly~~-convex lower-level with finite dimensional lower variables

$$\min_{\omega \in \mathbb{R}^d} \mathcal{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$

$$\text{s.t.} \quad \theta^\star_\omega \in \arg\min_{\theta \in \mathbb{R}^p} L_{in}(\omega, \theta)$$

Inner variable θ often represents the parameters of some predictive model h_θ, ex:

$$L_{in}(\omega, \theta) = \mathbb{E}\left[\|y - f_\theta(x)\|^2 + e^\omega \|f_\theta(x)\|^2\right]$$

- More sophisticated models f_θ, e.g. neural network $\implies \theta \mapsto L_{in}(\omega, \theta)$ is **non-convex.**
- The bad news: IFT not applicable for general non-convex losses: Non-uniqueness of $\theta^\star_\omega$.
- The worse news: The whole bilevel problem is ambiguous for non-convex losses:

  As many bilevel problems as choices of solution $\theta^\star_\omega$!

Non

~~Strongly~~-convex lower-level with finite dimensional lower variables

Towards non-convex implicit differentiation ([Arbel and Mairal, 2022] @ NeurIPS 2022)

- **Selection map** $\phi(\omega, \theta_0)$ as a replacement for ambiguous solution $\theta^{\star}_{\omega}$.
- Can be defined implicitly by an algorithmic procedure (**Implicit bias**),
  (e.g.: limit of gradient descent on $\theta \mapsto L_{in}(\omega, \theta)$ starting from initial location θ_0.)
- **Implicit differentiation** formula for $\nabla \mathscr{L}(\omega)$ still holds for suitable class of functions:

  Parametric Morse-Bott functions

Justifies using standard BO algorithms even when inner losses is non-convex!

Non ~~Strongly~~-convex lower-level with finite dimensional lower variables

Towards non-convex implicit differentiation ([Arbel and Mairal, 2022] @ NeurIPS 2022)

- **Selection map** $\phi(\omega,\theta_0)$ as a replacement for ambiguous solution $\theta^{\star}_{\omega}$.
- Can be defined implicitly by an algorithmic procedure (**Implicit bias**),
  (e.g.: limit of gradient descent on $\theta \mapsto L_{in}(\omega,\theta)$ starting from initial location θ_0.)
- **Implicit differentiation** formula for $\nabla\mathscr{L}(\omega)$ still holds for suitable class of functions:

Parametric Morse-Bott functions

Justifies using standard BO algorithms even when inner losses is non-convex!

Limitations

× Can still get instabilities in practice: ill-conditioned linear systems.

× Convergence analysis seems currently beyond reach.

Strongly-convex lower-level with infinite dimensional lower variables

Stay strongly-convex!

- Strong convexity is important for stability (both in theory and in practice).
- Allows precise control for the inner-level solution.

Go infinite dimensional!

- Increased expressivity: Beyond linear models.
- Opens way for theoretical analysis

Outline

Motivation: Objectives and challenges in bilevel optimization

Part I: Functional bilevel optimization

Part II: Towards a learning theory for Kernel Bilevel optimization

Functional bilevel optimization (NeurIPS 2024)

Functional Bilevel Optimization for Machine Learning

Ieva Petrulionyte, Julien Mairal, Michael Arbel
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
`firstname.lastname@inria.fr`

Abstract functional bilevel optimization

A Hidden functional structure in some BO

$$\min_{\omega\in\mathbb{R}^d} \mathscr{L}(\omega) := L_{out}(\omega, \theta^\star_\omega)$$
$$\text{s.t.} \quad \theta^\star_\omega \in \arg\min_{\theta\in\mathbb{R}^p} L_{in}(\omega,\theta)$$

Variable θ indexes a predictive model f_θ, ex:

$$L_{out}(\omega,\theta) = \mathbb{E}\left[\|y - f_\theta(x)\|^2\right]$$
$$L_{in}(\omega,\theta) = \mathbb{E}\left[\|y - f_\theta(x)\|^2 + e^\omega \|f_\theta(x)\|^2\right]$$

- Only model predictions $f_\theta(x)$ actually matter to both losses, not the parameters!
- Could choose a different parameterization without changing predictions.

Why not using f_θ as the inner variable instead of θ?

Abstract functional bilevel optimization

Leveraging the functional structure in BO

$$\min_{\omega \in \mathbb{R}^d} \mathscr{L}(\omega) := L_{out}(\omega, h^\star_\omega)$$
$$\text{s.t.} \quad h^\star_\omega \in \arg\min_{h \in \mathscr{H}} L_{in}(\omega, h)$$

Inner variable h is in a Hilbert space $\mathscr{H}$, ex:

$$L_{out}(\omega, h) = \mathbb{E}\left[\|y - h(x)\|^2\right]$$
$$L_{in}(\omega, h) = \mathbb{E}\left[\|y - h(x)\|^2 + e^{\omega}\|h(x)\|^2\right]$$

- **More flexible predictions**: Inner variable is a function in a rich Hilbert space $\mathscr{H}$.
- **Strong convexity**: easier to obtain in function spaces:
  Many ML objectives are (strongly) convex there (e.g.: MSE).
- **Function approximation**: Can approximate $h^\star_\omega$ using a model f_θ with parameters θ).
- **Implicit differentiation** w.r.t. ω performed *directly* on the function $h^\star_\omega$
  (not on parameters θ of approximating model f_θ!)

Implicit differentiation in an abstract Hilbert space $\mathscr{H}$

Theorem (informal): Assume that:

- There exists $\mu > 0$ such that $h \mapsto L_{in}(\omega', h)$ is μ-strongly convex for any $\omega \in \mathbb{R}^d$.
- L_{in} and L_{out} have finite values and are Fréchet (strongly) differentiable.
- $\partial_2 L_{in}$ is **Hadamard differentiable** (not necessarily Fréchet differentiable) .

Then, the total objective $\mathscr{L}$ is differentiable with gradient given by:

$$\nabla \mathscr{L}(\omega) = \partial_1 L_{out}(\omega, h_\omega^\star) + \nabla h_\omega^\star \partial_2 L_{out}(\omega, h_\omega^\star),$$

where the Jacobian $\nabla h_\omega^\star$ is the unique solution to the infinite dimensional system:

$$\nabla h_\omega^\star \overbrace{\partial_{2,2}^2 L_{in}(\omega, h_\omega^\star)}^{\text{Operator in } \mathscr{H}} = - \overbrace{\partial_{1,2}^2 L_{in}(\omega, h\omega^\star)}^{\text{Operator from } \mathscr{H} \text{ to } \mathbb{R}^d}$$

Implicit differentiation in an abstract Hilbert space $\mathscr{H}$

Theorem (informal): Assume that:

- There exists $\mu>0$ such that $h \mapsto L_{in}(\omega', h)$ is μ-strongly convex for any $\omega \in \mathbb{R}^d$.
- L_{in} and L_{out} have finite values and are Fréchet (strongly) differentiable.
- $\partial_2 L_{in}$ is **Hadamard differentiable** (not necessarily Fréchet differentiable) .

Then, the total objective $\mathscr{L}$ is differentiable with gradient given by:

$$\nabla \mathscr{L}(\omega) = \partial_1 L_{out}(\omega, h^\star_\omega) + \nabla h^\star_\omega \partial_2 L_{out}(\omega, h^\star_\omega),$$

where the Jacobian $\nabla h^\star_\omega$ is the unique solution to the infinite dimensional system:

$$\nabla h^\star_\omega \overbrace{\partial^2_{2,2} L_{in}(\omega, h^\star_\omega)}^{\text{Operator in } \mathscr{H}} = - \overbrace{\partial^2_{1,2} L_{in}(\omega, h\omega^\star)}^{\text{Operator from } \mathscr{H} \text{ to } \mathbb{R}^d}$$

- Standard versions of IFT require $\partial_2 L_{in}$ to be **Fréchet differentiable** → **Too restrictive**
  For L_2 spaces, [Nemirovski and Semenov, 1973] show it **only holds quadratic functions**.
- **Hadamard differentiability** allows a **broader** class of functions!
  (also used in statistics for the functional delta-method [van der Vaart and Wellner, 1996].)

Implicit differentiation in an abstract Hilbert space $\mathscr{H}$

Theorem (informal): Assume that:

- There exists $\mu > 0$ such that $h \mapsto L_{in}(\omega', h)$ is μ-strongly convex for any $\omega \in \mathbb{R}^d$.
- L_{in} and L_{out} have finite values and are Fréchet (strongly) differentiable.
- $\partial_2 L_{in}$ is **Hadamard differentiable** (not necessarily Fréchet differentiable) .

Then, the total objective $\mathscr{L}$ is differentiable with gradient given by:

$$\nabla \mathscr{L}(\omega) = \partial_1 L_{out}(\omega, h^\star_\omega) + \nabla h^\star_\omega \partial_2 L_{out}(\omega, h^\star_\omega),$$

where the Jacobian $\nabla h^\star_\omega$ is the unique solution to the infinite dimensional system:

$$\nabla h^\star_\omega \overbrace{\partial^2_{2,2} L_{in}(\omega, h^\star_\omega)}^{\text{Operator in } \mathscr{H}} = - \overbrace{\partial^2_{1,2} L_{in}(\omega, h\omega^\star)}^{\text{Operator from } \mathscr{H} \text{ to } \mathbb{R}^d}$$

✓ Expression of the gradient is **independent** of the way $h^\star_\omega$ is approximated.

× Abstract expression, unclear how to use it in practice.

Need to consider more concrete spaces!

Implicit differentiation in L_2 spaces and adjoint sensitivity method

A notable setting in ML (e.g.: Model-based RL, Instrumental Variables regression)

- Outer and inner losses are expectations of point-wise losses under data distribution $\mathbb{P}$:

$$L_{out}(\omega, h) := \mathbb{E}_{\mathbb{P}}\left[\ell_{out}(\omega, h(x), y)\right], \qquad L_{in}(\omega, h) := \mathbb{E}_{\mathbb{P}}\left[\ell_{in}(\omega, h(x), y)\right]$$

- $\mathcal{H} = L_2(\mathbb{P})$: The space of functions of x that are square integrable w.r.t. $\mathbb{P}$.

Proposition (informal): If $v \mapsto \ell_{in}(\omega, v, y)$ is strongly convex + mild assumptions:

$$\nabla \mathcal{L}(\omega) = \mathbb{E}_{\mathbb{Q}}\left[\partial_1 \ell_{out}\left(\omega, h^\star_\omega(x), y\right) + \partial_{1,2} \ell_{in}\left(\omega, h^\star_\omega(x), y\right) a^\star_\omega(x)\right],$$

where the adjoint function $a^\star_\omega$ is the unique minimizer in $L_2(\mathbb{P})$ of $a \mapsto L_{adj}(\omega, h^\star_\omega, a)$:

$$L_{adj}(\omega, h, a) := \mathbb{E}_{\mathbb{P}}\left[\frac{1}{2}\, a(x)^\top \partial^2_{2,2} \ell_{in}\left(\omega, h(x), y\right) a(x) + a(x)^\top \partial_2 \ell_{out}\left(\omega, h(x), y\right)\right]$$

Functional Bilevel Optmization (FuncBO)

General recipe

1. Approximate the search space for prediction and adjoint functions by flexible parametric families $(f_\theta)_{\theta\in\mathbb{R}^p}$ and $(g_\xi)_{\xi\in\mathbb{R}^q}$ (e.g. neural networks).
2. Optimize $\theta \mapsto L_{in}(\omega, f_\theta)$ using SGD (or any other algorithm) $\rightarrow f_\theta \approx h^\star_\omega$
3. Optimize $\xi \mapsto L_{adj}(\omega, f_\theta, g_\xi)$ using SGD (or any other optimizer) $\rightarrow g_\xi \approx a^\star_\omega$.
4. Approximate the gradient using a batch $\mathscr{B}$ of samples:

$$\widehat{\nabla\mathscr{L}}(\omega) := \frac{1}{|\mathscr{B}|}\sum_{(x,y)\in\mathscr{B}} \partial_1 \ell_{out}\left(\omega, f_\theta(x), y\right) + \partial_{1,2}\ell_{in}\left(\omega, f_\theta(x), y\right) g_\xi(x),$$

Functional Bilevel Optmization (FuncBO)

General recipe

1. Approximate the search space for prediction and adjoint functions by flexible parametric families $(f_\theta)_{\theta\in\mathbb{R}^p}$ and $(g_\xi)_{\xi\in\mathbb{R}^q}$ (e.g. neural networks).
2. Optimize $\theta \mapsto L_{in}(\omega, f_\theta)$ using SGD (or any other algorithm) $\rightarrow f_\theta \approx h_\omega^\star$
3. Optimize $\xi \mapsto L_{adj}(\omega, f_\theta, g_\xi)$ using SGD (or any other optimizer) $\rightarrow g_\xi \approx a_\omega^\star$.
4. Approximate the gradient using a batch $\mathscr{B}$ of samples:

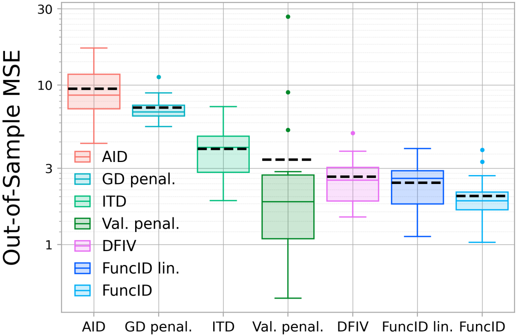
$$\widehat{\nabla\mathscr{L}}(\omega) := \frac{1}{|\mathscr{B}|}\sum_{(x,y)\in\mathscr{B}} \partial_1 \ell_{out}\left(\omega, f_\theta(x), y\right) + \partial_{1,2}\ell_{in}\left(\omega, f_\theta(x), y\right) g_\xi(x),$$

Advantages

- **Memory and time savings:** No need for higher-order derivatives of f_θ and g_ξ.
- **Stability:** Strongly-convex adjoint objective in function space: well-defined solution.

Applications to Instrumental Variables regression

Results on synthetic IV task [Xu et al., 2020]

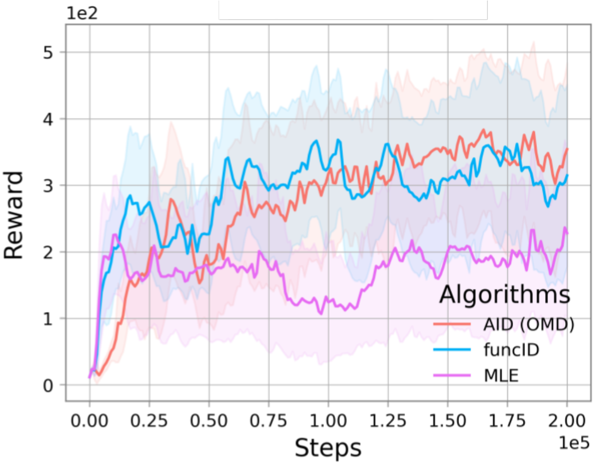


- IV solves a nested regression: **Functional bilevel problem!**
- Regressor functions are be deep networks: (DFIV [Xu et al., 2020])

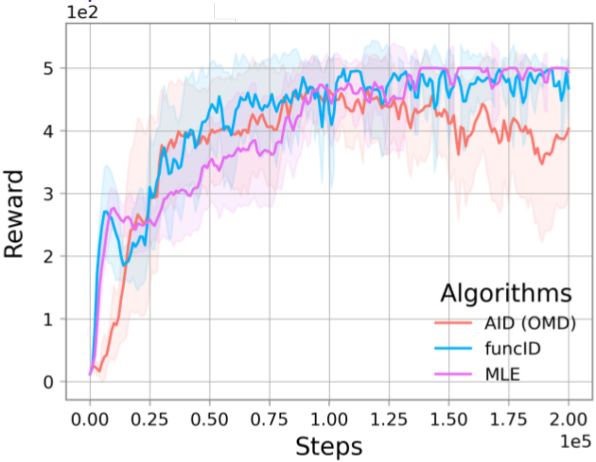
- Large improvement over classical bilevel optimization algorithms (AID/ITD)
- Competitive with problem specific methods (DFIV)!

Applications to Model-based RL (inspired by [Nikishin et al., 2022])

Miss-specified model



Well-specified model



Same method works well in both settings!

Convergence/generalization guarantees for FuncBO

Proposition (informal): Assume that $\mathscr{L}$ is L-smooth and admits a finite lower bound $\mathscr{F}^\star$. Consider an update rule $\omega_{t+1} = \omega_t - \eta \widehat{\nabla \mathscr{L}}(\omega_t)$ with suitable step-size η. Under mild smoothness assumptions ℓ_{in} and ℓ_{out}:

$$\min_{0 \le i \le t} \mathbb{E}\left[\left\| \nabla \mathscr{L}(\omega_i) \right\|^2 \right] \le \frac{4\left(\mathscr{F}(\omega_0) - \mathscr{F}^\star\right)}{\eta(t+1)} + 2\eta L \sigma_{eff}^2 + (c_1 \epsilon_{in} + c_2 \epsilon_{adj}),$$

where c_1, c_2, σ_{eff}^2 are positive constants, and $\epsilon_{in}, \epsilon_{adj}$ are sub-optimality errors that result from the inner and adjoint optimization procedures.

Convergence/generalization guarantees for FuncBO

Proposition (informal): Assume that $\mathscr{L}$ is L-smooth and admits a finite lower bound $\mathscr{F}^\star$. Consider an update rule $\omega_{t+1} = \omega_t - \eta \widehat{\nabla \mathscr{L}}(\omega_t)$ with suitable step-size η. Under mild smoothness assumptions ℓ_{in} and ℓ_{out}:

$$\min_{0 \le i \le t} \mathbb{E}\left[\left\|\nabla \mathscr{L}(\omega_i)\right\|^2\right] \le \frac{4\left(\mathscr{F}(\omega_0) - \mathscr{F}^\star\right)}{\eta(t+1)} + 2\eta L \sigma_{eff}^2 + (c_1 \epsilon_{in} + c_2 \epsilon_{adj}),$$

where c_1, c_2, σ_{eff}^2 are positive constants, and $\epsilon_{in}, \epsilon_{adj}$ are sub-optimality errors that result from the inner and adjoint optimization procedures.

- Sub-optimality errors are hard to control for neural networks:
  depends on optimization, approximation power, complexity of the class.
- Rates does not quantify the effect of sample size: (all hidden in the sub-optimality error).

Hard to get quantitative convergence results in L_2-spaces, what about other spaces?

Outline

Motivation: Objectives and challenges in bilevel optimization

Part I: Functional bilevel optimization

Part II: Towards a learning theory for Kernel Bilevel optimization

Kernel bilevel optimization: (work in progress)

Learning Theory for Kernel Bilevel Optimization

Fares El Khoury [1] **Edouard Pauwels** [2] **Samuel Vaiter** [3] **Michael Arbel** [1]

Reproducing Kernel Hilbert Spaces meet Bilevel optimization

Kernel Bilevel optimization

$$\min_{\omega \in \mathbb{R}^d} \mathscr{L}(\omega) := L_{out}(\omega, h^\star_\omega)$$
$$\text{s.t.} \quad h^\star_\omega \in \arg\min_{h \in \mathscr{H}} L_{in}(\omega, h)$$

$$L_{out}(\omega, h) = \mathbb{E}_{\mathbb{P}}\left[\ell_{out}(\omega, h(x), y)\right],$$
$$L_{in}(\omega, h) = \mathbb{E}_{\mathbb{P}}\left[\ell_{in}(\omega, h(x), y)\right] + \frac{\lambda}{2}\|h\|^2_{\mathscr{H}},$$

- Same as before, but now $\mathscr{H}$ is an RKHS with r.k. K.
- Already appeared in the past: [Keerthi et al., 2006, Kunapuli et al., 2008].

Reproducing Kernel Hilbert Spaces meet Bilevel optimization

Kernel Bilevel optimization

$$\min_{\omega \in \mathbb{R}^d} \mathcal{L}(\omega) := L_{out}(\omega, h^\star_\omega)$$
$$\text{s.t.} \quad h^\star_\omega \in \arg\min_{h \in \mathcal{H}} L_{in}(\omega, h)$$

$$L_{out}(\omega, h) = \mathbb{E}_{\mathbb{P}}\left[\ell_{out}(\omega, h(x), y)\right],$$
$$L_{in}(\omega, h) = \mathbb{E}_{\mathbb{P}}\left[\ell_{in}(\omega, h(x), y)\right] + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}},$$

- Same as before, but now $\mathcal{H}$ is an RKHS with r.k. K.
- Already appeared in the past: [Keerthi et al., 2006, Kunapuli et al., 2008].

Why an RKHS?

- **Expressiveness**: Some RKHS are dense in L_2 spaces [Steinwart and Christmann, 2008].
- **Algorithms**: Easy to derive thanks to the *Representer theorem* (next slide).
- **Learning theory:** Well-established theoretical framework available for regression (and similar problems) [Smale et al., 2005, Caponnetto and De Vito, 2007].

Practical algorithms for KBO

From infinite to finite dimensional bilevel optimization

- Have n i.i.d. samples (x_i, y_i).
- Replace expectations by empirical averages.
- **Representer theorem**: Optimal solution of the form:

$$\hat{h}_\omega = \sum_{i=1}^{n} (\hat{\theta}_\omega)_i K(x_i, .).$$

$$\min_{\omega \in \mathbb{R}^d} \widehat{\mathscr{L}}(\omega) := \frac{1}{n} \sum_{j=1}^{n} \ell_{out}(\omega, \left(\mathbf{K}\hat{\theta}_\omega\right)_j, y_j)$$

$$\text{s.t. } \hat{\theta}_\omega = \arg\min_{\theta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell_{in}(\omega, (\mathbf{K}\theta)_i, y_i) + \frac{\lambda}{2} \theta^\top \mathbf{K} \theta$$

Practical algorithms for KBO

From infinite to finite dimensional bilevel optimization

- Have n i.i.d. samples (x_i, y_i).
- Replace expectations by empirical averages.
- **Representer theorem**: Optimal solution of the form:

$$\hat{h}_\omega = \sum_{i=1}^{n} (\hat{\theta}_\omega)_i K(x_i, .).$$

$$\min_{\omega \in \mathbb{R}^d} \widehat{\mathscr{L}}(\omega) := \frac{1}{n} \sum_{j=1}^{n} \ell_{out}(\omega, (\mathbf{K}\hat{\theta}_\omega)_j, y_j)$$

$$\text{s.t. } \hat{\theta}_\omega = \arg\min_{\theta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell_{in}(\omega, (\mathbf{K}\theta)_i, y_i) + \frac{\lambda}{2} \theta^\top \mathbf{K} \theta$$

Optimizing $\widehat{\mathscr{L}}(\omega)$ by implicit differentiation

- Can use any standard bilevel optimization algorithm (AID [Lorraine et al., 2020], ITD [Franceschi et al., 2018]).

⚠ Dimension of θ increases with sample size: scalability issues, but not only ...

$\nabla \widehat{\mathscr{L}}(\omega)$ involves the Jacobian:

$$\nabla \hat{\theta}_\omega = -D_{1,2}^{in} \left(\mathbf{K} D_{2,2}^{in} + n\lambda \mathbb{1} \right)^{-1} \in \mathbb{R}^{d \times n},$$

$D_{1,2}^{in}$ and $D_{2,2}^{in}$: partial derivatives of ℓ_{in}.

Learning theory for KBO: Challenges

Limitation of existing generalization results for BO

Existing generalization results require inner-level parameters of **fixed dimension** [Bao et al., 2021, Zhang et al., 2024, Arbel and Mairal, 2021]:

- $\hat{\theta}_\omega$ and $\nabla\hat{\theta}_\omega$ approach *population counterparts* $\theta^\star_\omega \in \mathbb{R}^p$ and $\nabla\theta^\star_\omega \in \mathbb{R}^{d\times p}$ at rate $O(\frac{1}{\sqrt{n}})$.
- **Generalization results** obtained by controlling the error between true and estimated gradient $\|\nabla\mathcal{L}(\omega) - \nabla\widehat{\mathcal{L}}(\omega)\|$ in terms of $\|\hat{\theta}_\omega - \theta^\star_\omega\|$ and $\|\nabla\hat{\theta}_\omega - \nabla\theta^\star_\omega\|$.

Learning theory for KBO: Challenges

Limitation of existing generalization results for BO

Existing generalization results require inner-level parameters of **fixed dimension** [Bao et al., 2021, Zhang et al., 2024, Arbel and Mairal, 2021]:

- $\hat{\theta}_\omega$ and $\nabla\hat{\theta}_\omega$ approach *population counterparts* $\theta^\star_\omega \in \mathbb{R}^p$ and $\nabla\theta^\star_\omega \in \mathbb{R}^{d\times p}$ at rate $O(\frac{1}{\sqrt{n}})$.
- **Generalization results** obtained by controlling the error between true and estimated gradient $\|\nabla\mathscr{L}(\omega) - \nabla\widehat{\mathscr{L}}(\omega)\|$ in terms of $\|\hat{\theta}_\omega - \theta^\star_\omega\|$ and $\|\nabla\hat{\theta}_\omega - \nabla\theta^\star_\omega\|$.

Challenges with KBO

In KBO, the dimension of the parameter $\hat{\theta}_\omega$ grows with n:

- Expression of $\nabla\widehat{\mathscr{L}}(\omega)$ depends explicitly on vectors that grow with sample size, e.g. $\hat{\theta}_\omega$.
- No notion of limiting vector $\theta^\star_\omega$ for $\hat{\theta}_\omega$ exits!

Learning theory for KBO: Challenges

Limitation of existing generalization results for BO

Existing generalization results require inner-level parameters of **fixed dimension** [Bao et al., 2021, Zhang et al., 2024, Arbel and Mairal, 2021]:

- $\hat{\theta}_\omega$ and $\nabla\hat{\theta}_\omega$ approach *population counterparts* $\theta^\star_\omega \in \mathbb{R}^p$ and $\nabla\theta^\star_\omega \in \mathbb{R}^{d\times p}$ at rate $O(\frac{1}{\sqrt{n}})$.
- **Generalization results** obtained by controlling the error between true and estimated gradient $\|\nabla\mathscr{L}(\omega) - \nabla\widehat{\mathscr{L}}(\omega)\|$ in terms of $\|\hat{\theta}_\omega - \theta^\star_\omega\|$ and $\|\nabla\hat{\theta}_\omega - \nabla\theta^\star_\omega\|$.

Challenges with KBO

In KBO, the dimension of the parameter $\hat{\theta}_\omega$ grows with n:

- Expression of $\nabla\widehat{\mathscr{L}}(\omega)$ depends explicitly on vectors that grow with sample size, e.g. $\hat{\theta}_\omega$.
- No notion of limiting vector $\theta^\star_\omega$ for $\hat{\theta}_\omega$ exits!

Existing convergence/generalization results are not applicable to KBO!

Can we get an expression for $\nabla\widehat{\mathscr{L}}(\omega)$ independent of $\hat{\theta}_\omega$ and $\partial_\omega\hat{\theta}_\omega$?

Learning theory for KBO: A functional perspective

Rethinking the expression of $\nabla\widehat{\mathscr{L}}(\omega)$

$$
\begin{tikzcd}
\mathscr{L}(\omega) \arrow[r, "\text{Plug-in estimation}"] & \widehat{\mathscr{L}}(\omega) \arrow[d, "\text{Implicit differentiation}"] \\
 & \nabla\widehat{\mathscr{L}}(\omega)
\end{tikzcd}
$$

1. Discretize the problem using samples,
2. Apply implicit differentiation.

Pros/cons

✓ Convenient in practice: can use classical BO algorithms!

× Hard to use for statistical theory.

What if we followed another path?

Learning theory for KBO: A functional perspective

An alternative (functional) expression for $\nabla\widehat{\mathscr{L}}(\omega)$

$$\begin{tikzcd}
\mathscr{L}(\omega) \arrow[rr, "\text{Plug-in estimation}"] \arrow[d, "\text{Implicit differentiation}"'] & & \widehat{\mathscr{L}}(\omega) \arrow[d, "\text{Implicit differentiation}"] \\
\nabla\mathscr{L}(\omega) \arrow[rr, "\text{Plug-in estimation}"'] & & \widehat{\nabla\mathscr{L}}(\omega)
\end{tikzcd}$$

1. Implicit differentiation in $\mathscr{H}$ to express $\nabla\mathscr{L}(\omega)$ in terms of inner solution $h_\omega^\star$ and adjoint $a_\omega^\star$.
2. Discretize expectations in $\nabla\mathscr{L}(\omega)$ + replace $h_\omega^\star$ and $a_\omega^\star$ by empirical estimates $\hat{h}_\omega$ and $\hat{a}_\omega$.

Pros/cons

- × Not very practical expression.
- ✓ Can control $\|\nabla\mathscr{L}(\omega) - \nabla\widehat{\mathscr{L}}(\omega)\|$ in terms of $\|\hat{h}_\omega - h_\omega^\star\|_{\mathscr{H}}$ and $\|\hat{a}_\omega - a_\omega^\star\|_{\mathscr{H}}$.

Can use tools from learning theory to control $\|\hat{h}_\omega - h_\omega^\star\|_{\mathscr{H}}$ and $\|\hat{a}_\omega - a_\omega^\star\|_{\mathscr{H}}$!

Learning theory for KBO: A functional perspective

An alternative (functional) expression for $\nabla \widehat{\mathscr{L}}(\omega)$

$$
\begin{tikzcd}
\mathscr{L}(\omega) \arrow[r, "\text{Plug-in estimation}"] \arrow[d, "\text{Implicit differentiation}"'] & \widehat{\mathscr{L}}(\omega) \arrow[d, "\text{Implicit differentiation}"] \\
\nabla \mathscr{L}(\omega) \arrow[r, "\text{Plug-in estimation}"'] & \nabla \widehat{\mathscr{L}}(\omega)
\end{tikzcd}
$$

Both paths yield the same estimator!

A learning theory for KBO: Main results

Maximal inequalities

Theorem (informal): Fix a compact subset Ω of $\mathbb{R}^d$. Then, under mild assumptions:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathscr{L}(\omega)-\widehat{\nabla\mathscr{L}}(\omega)\right\|\right]\lesssim\frac{1}{\sqrt{n}},$$

A learning theory for KBO: Main results

Maximal inequalities

Theorem (informal): Fix a compact subset Ω of $\mathbb{R}^d$. Then, under mild assumptions:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathscr{L}(\omega)-\widehat{\nabla\mathscr{L}}(\omega)\right\|\right]\lesssim\frac{1}{\sqrt{n}},$$

Generalization error of BO algorithms

Corollary (informal): Consider the iterates $\omega_{t+1}=\omega_t-\eta\nabla\widehat{\mathscr{L}}(\omega_t)$. Then, assuming ℓ_{out} is coercive in its second argument, it follows that:

$$\mathbb{E}\left[\min_{i=0,\dots,t}\left\|\nabla\mathscr{L}(\omega_i)\right\|\right]\lesssim\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{t+1}}.$$

A learning theory for KBO: Main results

Maximal inequalities

Theorem (informal): Fix a compact subset Ω of $\mathbb{R}^d$. Then, under mild assumptions:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathcal{L}(\omega)-\widehat{\nabla\mathcal{L}}(\omega)\right\|\right]\lesssim\frac{1}{\sqrt{n}},$$

Proof sketch and challenges

1 Upper-bound $\|\nabla\mathcal{L}(\omega)-\nabla\widehat{\mathcal{L}}(\omega)\|$ by terms:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tau_\omega(x_i,y_i)-\mathbb{E}_{\mathbb{P}}\left[\tau_\omega(x_i,y_i)\right]\right\|,$$

where τ_ω can be **real-valued** functions or **RKHS-valued** ones (e.g. values in $\mathcal{H}$).

A learning theory for KBO: Main results

Maximal inequalities

Theorem (informal): Fix a compact subset Ω of $\mathbb{R}^d$. Then, under mild assumptions:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathscr{L}(\omega)-\widehat{\nabla\mathscr{L}}(\omega)\right\|\right]\lesssim\frac{1}{\sqrt{n}},$$

Proof sketch and challenges

1. Upper-bound $\|\nabla\mathscr{L}(\omega)-\nabla\widehat{\mathscr{L}}(\omega)\|$ by terms:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tau_\omega(x_i,y_i)-\mathbb{E}_{\mathbb{P}}[\tau_\omega(x_i,y_i)]\right\|,$$

   where τ_ω can be **real-valued** functions or **RKHS-valued** ones (e.g. values in $\mathscr{H}$).

2. **Real-valued** τ_ω: Maximal inequalities for empirical processes [Van der Vaart, 2000].
3. **RKHS-valued** τ_ω: Maximal inequalities for U-processes [Sherman, 1994]. (Applied to the squared error!)

⚠ Degraded rates when using "simpler" approaches for RKHS-valued τ_ω.

Summary

A new framework for bilevel optimization in ML

- Compatible with flexible function approximation tools (neural networks, RKHS).
- Practical algorithms like FuncBO
- Opens way for generalization theory

Limitations and future work

- Extension to other spaces of functions: ex: Sobolev spaces: learning PDEs
- Generalization theory beyond RKHS: deep networks?

Summary

A new framework for bilevel optimization in ML

- Compatible with flexible function approximation tools (neural networks, RKHS).
- Practical algorithms like FuncBO
- Opens way for generalization theory

Limitations and future work

- Extension to other spaces of functions: ex: Sobolev spaces: learning PDEs
- Generalization theory beyond RKHS: deep networks?

Thank you!

Arbel, M. and Mairal, J. (2021).
Amortized implicit differentiation for stochastic bilevel optimization.
In International Conference on Learning Representations (ICLR) 2022.

Arbel, M. and Mairal, J. (2022).
Non-convex bilevel games with critical point selection maps.
Advances in Neural Information Processing Systems, 35:8013–8026.

Bao, F., Wu, G., Li, C., Zhu, J., and Zhang, B. (2021).
Stability and generalization of bilevel programming in hyperparameter optimization.
Advances in Neural Information Processing Systems, 34:4529–4541.

Bolte, J., Le, Q.-T., Pauwels, E., and Vaiter, S. (2024).
Geometric and computational hardness of bilevel programming.
arXiv preprint arXiv:2407.12372.

Caponnetto, A. and De Vito, E. (2007).
Optimal rates for the regularized least-squares algorithm.
Foundations of Computational Mathematics, 7(3):331–368.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. (2022).
A framework for bilevel optimization that enables stochastic and global variance reduction algorithms.
arXiv preprint arXiv:2201.13409.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. (2017).
Forward and reverse gradient-based hyperparameter optimization.
In International Conference on Machine Learning, pages 1165–1173. PMLR.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. (2018).
Bilevel programming for hyperparameter optimization and meta-learning.
In International Conference on Machine Learning, pages 1568–1577. PMLR.

Ghadimi, S. and Wang, M. (2018).
Approximation methods for bilevel programming.
arXiv preprint arXiv:1802.02246.

Ji, K., Yang, J., and Liang, Y. (2021).
Bilevel optimization: Convergence analysis and enhanced design.
In International Conference on Machine Learning, pages 4882–4892. PMLR.

Keerthi, S., Sindhwani, V., and Chapelle, O. (2006).
An Efficient Method for Gradient-Based Adaptation of Hyperparameters in SVM Models.

Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J.-S. (2008).
Classification model selection via bilevel programming.
Optimization Methods & Software, 23(4):475–489.

Lorraine, J., Vicol, P., and Duvenaud, D. (2020).
Optimizing millions of hyperparameters by implicit differentiation.
In International Conference on Artificial Intelligence and Statistics, pages 1540–1552. PMLR.

Mairal, J., Bach, F., and Ponce, J. (2011).
Task-driven dictionary learning.
IEEE transactions on pattern analysis and machine intelligence, 34(4):791–804.

Nemirovski, A. and Semenov, S. (1973).
On polynomial approximation of functions on hilbert space.
Mathematics of the USSR-Sbornik, 21(2):255.

Nikishin, E., Abachi, R., Agarwal, R., and Bacon, P.-L. (2022).

Control-oriented model-based reinforcement learning with implicit differentiation.
In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 7886–7894.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. (2019).
Meta-Learning with Implicit Gradients.
In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems 32 (NeurIPS). Curran Associates, Inc.

Sherman, R. P. (1994).
Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators.
The Annals of Statistics, 22(1):439 – 459.

Smale, S., Smale, S., and Zhou, D.-x. (2005).
Learning theory estimates via integral operators and their approximations.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer Publishing Company, Incorporated, 1st edition.

Van der Vaart, A. W. (2000).
Asymptotic statistics, volume 3.
Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996).
Weak Convergence, pages 16–28.
Springer New York, New York, NY.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. (2020).
Learning deep features in instrumental variable regression.
arXiv preprint arXiv:2010.07154.

Zhang, X., Chen, H., Gu, B., Gong, T., and Zheng, F. (2024).
Fine-grained analysis of stability and generalization for stochastic bilevel optimization.
In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, pages 5508–5516.

Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. (2022).
Revisiting and advancing fast adversarial training through the lens of bi-level optimization.
In International Conference on Machine Learning, pages 26693–26712. PMLR.