Learning Theory for Kernel Bilevel Optimization

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

[1]Inria & Univ. Grenoble Alpes [2]Toulouse School of Economics [3]CNRS & LJAD, Nice

NeurIPS 2025

Setup

Nonparametric Bilevel Optimization

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

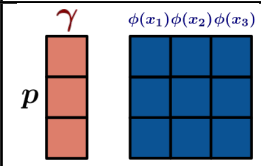
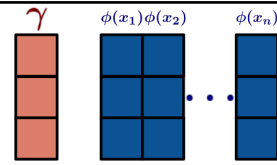
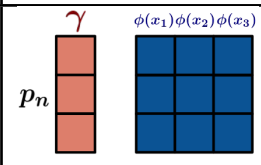
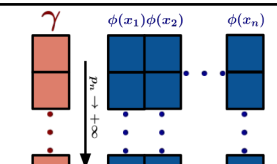
- $\mathscr{H}$: hypothesis space.
- $L_{out}(\omega, h) = \mathbb{E}_{\mathbb{Q}}[\ell_{out}(\omega, h(x), y)]$.
- $L_{in}(\omega, h) = \mathbb{E}_{\mathbb{P}}[\ell_{in}(\omega, h(x), y)] + \frac{\lambda}{2}\|h\|^2_{\mathscr{H}}$, $\lambda > 0$: regularization.

Goal

Establish a learning-theoretic foundation for nonparametric bilevel optimization.

Motivation

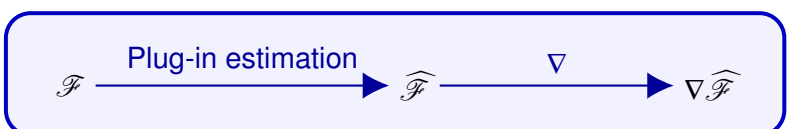
- $h^{\gamma}(x_i) = \gamma^{\top}\phi(x_i)$, with γ, $\phi(x_i) \in \mathbb{R}^p$ and $(x_i)_{1 \leq i \leq n}$ n i.i.d. samples.

	$n = 3$	$n \to +\infty$
Parametric • p fixed • Limited expressivity		
Nonparametric • $p \asymp n$ • Expressive models		

- Generalization in the nonparametric setting is **underexplored**.
- **This work**: $\mathscr{H}$ = RKHS (Kernel Bilevel Optimization (**KBO**)).

Practical Algorithms for KBO

- **Solving** KBO using gradient-based methods requires **computing** $\nabla\mathscr{F}(\omega)$.
- $\nabla\mathscr{F}(\omega)$ is **intractable** due to differentiation in an RKHS.
- An **estimator** of $\nabla\mathscr{F}(\omega)$ is needed.
- **Prior works**:



From Infinite to Finite Dimensional

- Have access to i.i.d. samples $(x_i, y_i)_{1\le i\le n} \sim \mathbb{P}$, $(\tilde{x}_j, \tilde{y}_j)_{1\le j\le m} \sim \mathbb{Q}$.
- **Representer theorem**: $\hat{h}_\omega(x) = \hat{\gamma}_\omega^\top \phi(x)$ (optimal solution of the discretized inner problem).

$$\min_{\omega\in\mathbb{R}^d} \widehat{\mathscr{F}}(\omega) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \ell_{out}(\omega, \hat{\gamma}_\omega^\top \phi(\tilde{x}_j), \tilde{y}_j) \qquad \text{(KBO)}$$

$$\text{s.t.} \quad \hat{\gamma}_\omega = \operatorname*{arg\,min}_{\gamma\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \ell_{in}(\omega, \gamma^\top \phi(x_i), y_i) + \frac{\lambda}{2}\gamma^\top \Phi \gamma,$$

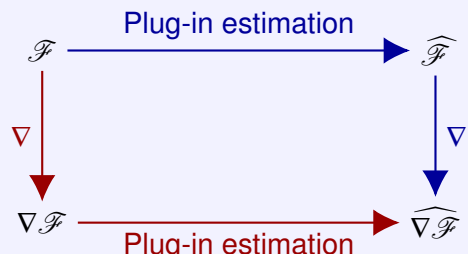
where $\Phi = [\phi(x_1), \ldots, \phi(x_n)] \in \mathbb{R}^{n\times n}$ is the kernel matrix.

- **No limiting vector** for $\hat{\gamma}_\omega$.
- Existing generalization results for bilevel optimization are **not applicable**.

Equivalence of Gradient Estimators

Proposition

The following diagram is **commutative**.



— Estimator used in practice.

— Estimator used in a statistical analysis.

Generalization Bounds for BO Algorithms

Corollary (Generalization for bilevel gradient descent)

Let $\omega_{t+1} = \omega_t - \eta \widehat{\nabla \mathscr{F}}(\omega_t)$. Under technical assumptions, we have:

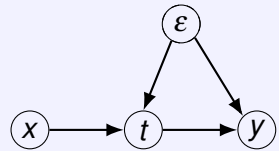
$$\mathbb{E}\Big[\min_{i=0,\dots,t} \|\nabla \mathscr{F}(\omega_i)\|\Big] \lesssim \underbrace{\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}}_{\text{statistical component}} + \underbrace{\frac{1}{\sqrt{t+1}}}_{\text{optimization component}}$$

- **Algorithmic insight**: convergence requires a **balance** between **data availability** and **computational budget**.

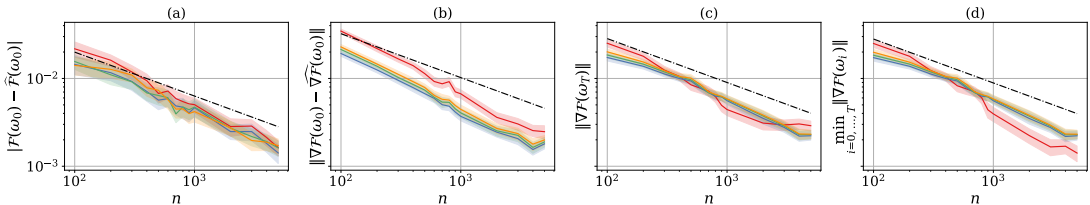
Experiments: Synthetic Instrumental Variable Regression

$$L_{out}(\omega, h) = \frac{1}{2}\mathbb{E}_{x,y}\left[|h(x) - y|^2\right].$$

$$L_{in}(\omega, h) = \frac{1}{2}\mathbb{E}_{x,t}\left[|h(x) - \omega^\top \Psi(t)|^2\right] + \frac{\lambda}{2}\|h\|^2_{\mathscr{H}}.$$



Structural causal model: $y = \omega^{\star\top}\Psi(t) + \varepsilon$. **Goal**: estimate $\omega^\star$.



More in the Paper

- Details on the numerical experiments.
- Maximal inequalities for KBO.
- Detailed proof using a novel technique involving U-processes.
- Generalization results for projected bilevel gradient descent.

Thank you for listening!