On Generalization in Bilevel Least Squares with Overparameterized Neural Networks

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 2026

Setup and Goal

Bilevel Least Squares

$$\min_{\omega \in \Omega} \mathscr{F}(\omega) := L_{out}(h^\star_\omega) \quad \text{such that} \quad h^\star_\omega = \arg\min_{h \in L^2} L_{in}(\omega, h).$$

- $L_{in}(\omega, h) = \frac{1}{2}\mathbb{E}_{(x,z)}\left[|h(x) - g_\omega(z)|^2\right], \quad L_{out}(h) = \frac{1}{2}\mathbb{E}_{(x,y)}\left[|h(x) - y|^2\right].$
- **Unique** inner solution: $h^\star_\omega(x) = \mathbb{E}[g_\omega(z) \mid x]$.
- **Applications**: instrumental variable regression, value function fitting in RL, etc.

Goal

Understand how a **gradient-based** bilevel algorithm **generalizes** when $h^\star_\omega$ is learned by **overparameterized neural networks**.

Hypergradient and Algorithm

Proposition (Hypergradient)

$\mathscr{F}$ is differentiable and $\mathscr{L}$-smooth, with

$$\nabla \mathscr{F}(\omega) = -\mathbb{E}_{(x,z)}\left[\partial_\omega g_\omega(z)\, a^\star_\omega(x)\right], \qquad a^\star_\omega = -h^\star_\omega + b^\star, \qquad b^\star(x) = \mathbb{E}[y \mid x].$$

Computing $\nabla \mathscr{F}$ **decomposes into two regressions**: $h^\star_\omega$ and $b^\star$.

Projected Mini-Batch Biased SGD

Require: $\mathscr{D}_{\text{in}} = \{(x_i, z_i)\}_{i=1}^{n_{\text{in}}}$, $\mathscr{D}_{\text{out}} = \{(x_j, y_j)\}_{j=1}^{n_{\text{out}}}$

1: Estimate $\hat{b}$ on $\mathscr{D}_{\text{out}}$ (**once**)
2: **for** $k = 0, \ldots, K-1$ **do**
3: Estimate $\hat{h}_{\omega_k}$ on $\mathscr{D}_{\text{in}}$ (**each outer iteration**)
4: Draw a fresh mini-batch and form the plug-in estimate $\widehat{\nabla \mathscr{F}}(\omega_k)$
5: $\omega_{k+1} \leftarrow \Pi_\Omega\left(\omega_k - \eta\, \widehat{\nabla \mathscr{F}}(\omega_k)\right)$

Convergence under Inexact Hypergradients

- **Gradient mapping**: $\mathscr{G}_\lambda(\omega) = \frac{1}{\lambda}\big(\omega - \Pi_\Omega(\omega - \lambda \nabla \mathscr{F}(\omega))\big)$, $= 0$ iff ω is **stationary**.

Theorem (Convergence of projected mini-batch biased SGD)

For $0 < \eta \leq \frac{1}{2\mathscr{L}}$:

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\big[\|\mathscr{G}_{1/\mathscr{L}}(\omega_k)\|^2\big] \lesssim \underbrace{\frac{1}{\eta K}}_{\text{optimization}} + \underbrace{\frac{\eta\,\sigma^2}{n_g}}_{\text{variance}} + \underbrace{e_h + e_b}_{\text{bias}},$$

with $e_h = \mathbb{E}\big[\sup_\omega \|\hat{h}_\omega - h^\star_\omega\|^2_{L^2}\big]$, $e_b = \mathbb{E}\big[\|\hat{b} - b^\star\|^2_{L^2}\big]$.

Question

Can we obtain **explicit rates** for e_h, e_b with **overparameterized neural networks**?

Explicit Rates in the NTK Regime

- $\hat{h}_\omega$, $\hat{b}$: **two-layer ReLU networks** $f_\theta(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \sigma(u_i^\top x + c_i) + a$, trained by **gradient flow** with **early stopping**, and **truncated** outside high-probability events.

Proposition (Generalization bounds in the NTK regime)

Let s_h, s_b be the **smoothness** of $h^\star_\omega$, $b^\star$, and $\gamma_h = \frac{p+1}{s_h(p+1)+p}$, $\gamma_b = \frac{p+1}{s_b(p+1)+p}$.
For sufficiently wide networks and suitable stopping times:

$$e_h \lesssim d^4 (\ln n_{\text{in}})^4 \, n_{\text{in}}^{-s_h \gamma_h}, \qquad e_b \lesssim (\ln n_{\text{out}})^4 \, n_{\text{out}}^{-s_b \gamma_b}.$$

Corollary (End-to-end sample complexity)

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\| \mathscr{G}_{1/\mathscr{L}}(\omega_k) \|^2 \right] \lesssim \frac{1}{\eta K} + \frac{\eta}{n_g} + d^4 (\ln n_{\text{in}})^4 \, n_{\text{in}}^{-s_h \gamma_h} + (\ln n_{\text{out}})^4 \, n_{\text{out}}^{-s_b \gamma_b}.$$

$\Rightarrow$ The **slower** rate dominates: **balance** n_{in} and n_{out}.

Experiment: Validation in the NTK Regime

- **Cusp targets** on $[0,1]^2$: $h^\star_\omega(x) = \sum_{j=1}^4 \omega_j \|x - c_j\|^{q_h}$, $b^\star(x) = \|x - x_0\|^{q_b}$.
- **Known smoothness** $\bar{s} = \frac{2q+2}{3} \in \{0.70, 0.83, 0.97\} \Rightarrow$ **explicit theoretical rates**.

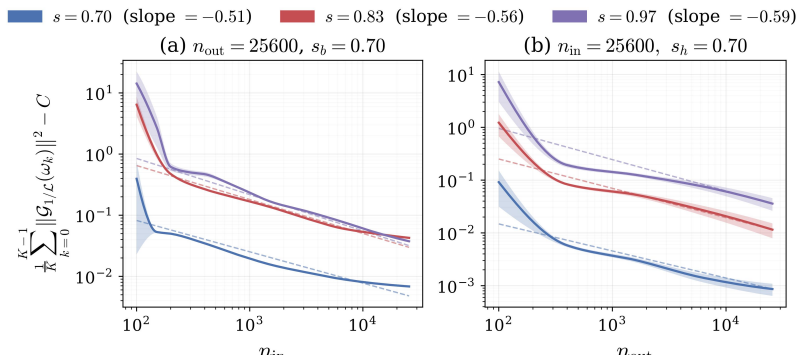


Figure: Near-stationarity vs. n_{in} (a) and n_{out} (b); dashed lines: theoretical slopes.

More in the Paper

- Full statements with explicit constants and assumptions.
- Construction of the truncation events.
- Detailed proofs.
- Sobolev characterization of the NTK interpolation spaces.
- Experiments **beyond the NTK regime**: an **implicit bias** of the bilevel structure toward feature learning.

Thank you for listening!