Structured Strategies for Nonsmooth Bilevel Hyperparameter Optimization

David Villacís

Universidad LOYOLA

joint work with Pedro Pérez-Aros and Emilio Vilches

SIAM Optimization 2026 — Edinburgh, UK

1. Motivation & Problem Statement

Context

Variational formulation of an inverse problem parametrized by $x \in X$:

$$\min_{y \in Y} F(\mathcal{A}(y); d) + G_x(y), \quad G_x(y) = \sum_{i=1}^{n} \psi(x_i)|y_i|.$$

- Weighted regularizers are valuable when the model needs to be **robust** and handle **small and structured disturbances**.
- Assigning different weights to individual features lets the model **adapt** flexibly to uncertainty.
- Maintains **stable** performance under noise or variation.

Xu et. al. (2008), Guo et. al. (2023)

Applications

Overparametrized regression

$$\min_{y \in \mathbb{R}^n} \frac{1}{2} \parallel A^{\text{train}} y - y^{\text{train}} \parallel^2 + \sum_{i=1}^{n} \exp(x_i)|y_i|$$

where $A^{\text{train}} \in \mathbb{R}^{n \times m}$ with $n \ll m$ and $y^{\text{train}} \in \mathbb{R}^n$ are the sample matrix and the corresponding target vector, respectively.

Applications

Wavelet-based image restoration

$$\min_{y \in \mathbb{R}^n} \frac{1}{2} \parallel RWy - y^{\text{damaged}} \parallel^2 + \sum_{i=1}^{n} \exp(x_i)|y_i|$$

where R is a blurring matrix and W contains a wavelet basis.

- Multiplying by W corresponds to performing the inverse wavelet transform.

Challenge

The hyperparameter x is **not known** and must be learned from data.

Grid search: evaluate the model for a set of hyperparameter values and select the one that minimizes the validation error. $\rightarrow$ Cross-validation.

Bayesian optimization: model the hyperparameter space with a probabilistic model and select the hyperparameter that maximizes the expected improvement. $\rightarrow$ low dimensionality.

2. Learning Model

Bilevel Hyperparameter Learning

We consider the optimal hyperparameter as the **solution of a bilevel optimization** problem:

$$\min_{x \in X} L(x, y^*(x)) \text{ s.t. } y^*(x) \in S(x) := \arg\min_{y \in Y} \{F(y) + G_x(y)\}$$

where L evaluates the performance of the learned hyperparameter on a **validation set**, and $S(x)$ denotes the solution set of the lower-level problem, obtained from the **training set**.

Smooth case & the gap

If S is **single-valued and C^1**, the bilevel problem is just $\min_{x \in X} \Phi(x) := L(x, S(x))$, with IFT hypergradient

$$\nabla_x \Phi(x) = \nabla_x L(x, S(x)) + S'(x)^\top \nabla_y L(x, S(x)).$$

Gap. Our G_x is **not differentiable**: S need not be differentiable or single-valued, and smoothing $|\cdot|$ shifts the minimizer off the ℓ_1 boundary, **biasing** the upper level.

3. On characterizing a hyper-subdifferential

Let us introduce the **forward-backward** operator and the **residual** operator as follows:

$$T^\gamma(x,y) := \operatorname{prox}_{\gamma G_x}(y - \gamma \nabla F(y)) = (\mathcal{T} \circ \mathcal{H}^\gamma)(x,y),$$
$$R^\gamma(x,y) := \gamma^{-1}(y - T^\gamma(x,y)),$$

where $\mathcal{T}$ is the **component-wise soft-thresholding** operator

$$\mathcal{T}(u,v) := (\operatorname{sign}(v_i)\max(|v_i| - \gamma, 0))_{i=1}^n,$$
$$\mathcal{H}^\gamma(x,y) := (\gamma\Psi(x), y - \gamma\nabla F(y)).$$

Bilevel FB reformulation

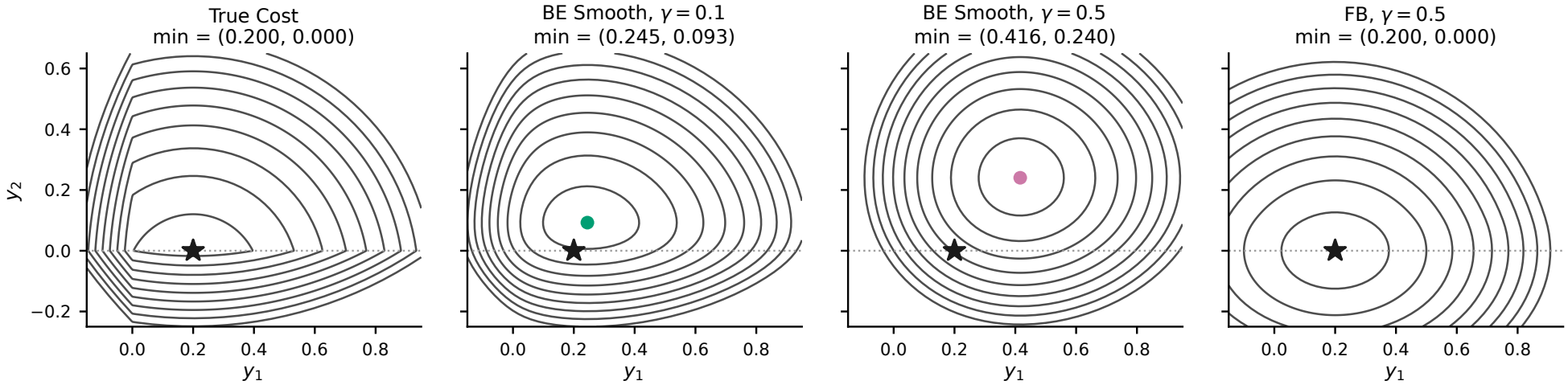
$$\min_{x \in X} L(x, y^*(x)) \quad \text{s.t.} \quad y^*(x) \in S(x) := \{y \in Y \mid R^\gamma(x, y) = 0\}.$$

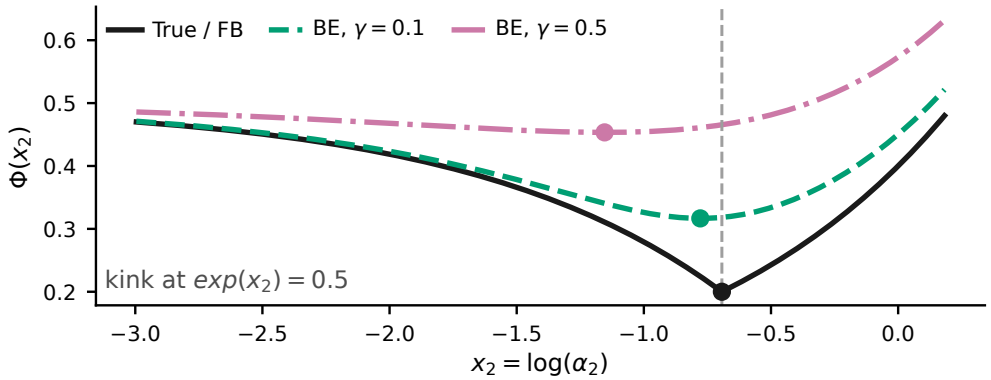
Proposition: Let $\gamma > 0$. For every $x \in X$ and $y \in Y$,

$$R^\gamma(x, y) = 0 \iff 0 \in \nabla F(y) + \partial G_x(y) \iff y \in S(x).$$

$\Longrightarrow$ S is **single-valued** and **locally Lipschitz continuous** on X.

Bilevel FB reformulation





Bilevel FB reformulation

Strict differentiability of L + FB reformulation give

$$\partial\Phi(\bar{x}) = \nabla_x L(\bar{x},\bar{y}) + D^*S(\bar{x})\big(\nabla_y L(\bar{x},\bar{y})\big);$$

the missing piece D^*S comes from R^γ.

Lemma: Let $\gamma > 0$, $\bar{x} \in X$, $\bar{y} := S(\bar{x})$. If $\ker\, D^*R^\gamma(\bar{x},\bar{y}) = \{0\}$, then

$$D^*S(\bar{x})(y^*) \subseteq \{x^* \mid (x^*,-y^*) \in D^*R^\gamma(\bar{x},\bar{y})(z^*), \exists z^* \in Y\},$$

with **equality** when R^γ is **lower-regular** at $(\bar{x},\bar{y})$.

Residual-enlarged subdifferential

The chain-rule inclusion for D^*S via D^*R^γ may be **strict** — we name the right-hand side as a tractable **outer approximation**.

Definition: The **residual-enlarged coderivative** of S at $\bar{x}$ for $y^* \in Y$:

$$D^*_{R^\gamma} S(\bar{x})(y^*) := \{x^* \in X \mid (x^*, -y^*) \in D^*R^\gamma(\bar{x}, \bar{y})(z^*), \exists z^* \in Y\},$$

and the **residual-enlarged subdifferential** of Φ at $\bar{x}$:

$$\partial^\gamma_R \Phi(\bar{x}) := \nabla_x L(\bar{x}, \bar{y}) + D^*_{R^\gamma} S(\bar{x})\big(\nabla_y L(\bar{x}, \bar{y})\big).$$

Then $\partial\Phi(\bar{x}) \subseteq \partial^\gamma_R \Phi(\bar{x})$, with **equality** when R^γ is **lower-regular** at $(\bar{x}, \bar{y})$.

4. On computing a hyper-subgradient

Adjoint system

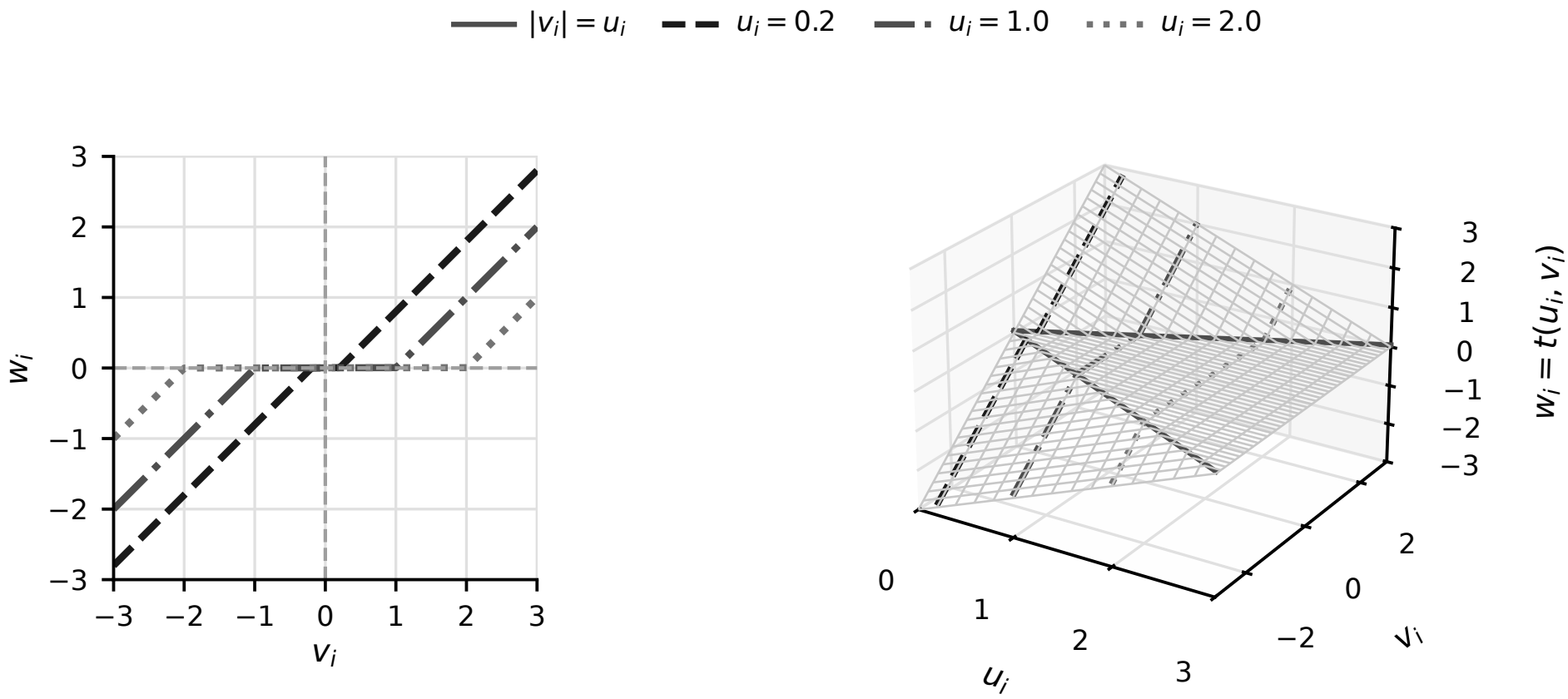
At the prox argument $(\bar{u}, \bar{v}) := \mathcal{H}^\gamma(\bar{x}, \bar{y})$, the chain reduces **componentwise** to the **soft-thresholding coderivative**. For $z^* \in \mathbb{R}^p$, $D^*\mathcal{T}(\bar{u}, \bar{v})(z^*)$ collects all (u^*, v^*) with

$$\begin{cases} (u_i^*, v_i^*) = (-z_i^*, z_i^*) & \text{if } \; i \in \mathcal{I}^+ \quad (\bar{v}_i > \bar{u}_i) \\ (u_i^*, v_i^*) = (z_i^*, z_i^*) & \text{if } \; i \in \mathcal{I}^- \quad (\bar{v}_i < -\bar{u}_i) \\ (u_i^*, v_i^*) = (0, 0) & \text{if } \; i \in \mathcal{A} \quad (|\bar{v}_i| < \bar{u}_i) \\ \text{three options} \quad (\text{multi-valued}) & \text{if } \; i \in \mathcal{B}^\pm \quad (\bar{v}_i = \pm \bar{u}_i). \end{cases}$$

Lemma: (Adjoint system, Lemma 12.) Assuming $\ker D^* R^\gamma(\bar{x}, \bar{y}) = \{0\}$, $x^* \in D^*_{R^\gamma} S(\bar{x})(z^*)$ iff there exist $q \in \mathbb{R}^p$ and $(u^*, v^*) \in D^*\mathcal{T}(\mathcal{H}^\gamma(\bar{x}, \bar{y}))(-q)$ with

$$-z^* = q + \big(I - \gamma \nabla^2 F(\bar{y})\big) v^*, \quad x^* = \gamma J \Psi(\bar{x}) u^*.$$

Soft-thresholding coderivative



Soft-thresholding coderivative

Componentwise, $D^*\mathcal{T}(\bar{u},\bar{v})(z^*)$ collects all (u^*,v^*) such that, for each i,

$$\begin{cases} (u_i^*, v_i^*) = (-z_i^*, z_i^*) & \text{if } \ i \in \mathcal{I}^+ \\ (u_i^*, v_i^*) = (z_i^*, z_i^*) & \text{if } \ i \in \mathcal{I}^- \\ u_i^* = v_i^* = 0 & \text{if } \ i \in \mathcal{A} \\ u_i^* = v_i^* = 0 \quad \vee & \\ u_i^* = -v_i^*, v_i^* \in [0, z_i^*] \quad \vee & \text{if } \ i \in \mathcal{B}^+ \\ u_i^* = -z_i^*, v_i^* = z_i^* & \\ u_i^* = v_i^* = 0 \quad \vee & \\ u_i^* = v_i^*, v_i^* \in [z_i^*, 0] \quad \vee & \text{if } \ i \in \mathcal{B}^-. \\ u_i^* = z_i^*, v_i^* = z_i^* & \end{cases}$$

A diagonal **selector** $D_{\text{supp}} \in \{-1, 0, +1\}^{p\times p}$ encodes the chosen (u^*, v^*) per coordinate: **fixed** at ± 1 on $\mathcal{I}^{\pm}$ and 0 on $\mathcal{A}$; **free** on $\mathcal{B}^{\pm}$ — the **oracle freedom**.

Sign-consistent biactive selection (SC)

Want $\sigma_i q_i > 0$ for every selected biactive i — the adjoint must **corroborate** the sign it was assigned.

Definition: Let $z^* := \nabla_y L(\bar{x}, \bar{y})$ and $H_{\mathcal{S}} := \gamma[\nabla^2 F(\bar{y})]_{\mathcal{S},\mathcal{S}}$.

1. **Init.** $\sigma_i^{(0)} = +1$ for $i \in \mathcal{I}^+ \cup \{j \in \mathcal{B}^+ : z_j^* < 0\}$; $\sigma_i^{(0)} = -1$ for $i \in \mathcal{I}^- \cup \{j \in \mathcal{B}^- : z_j^* > 0\}$; $\sigma_i^{(0)} = 0$ otherwise. Set $\mathcal{S}^{(0)} := \{i : \sigma_i^{(0)} \neq 0\}$.
2. **Prune.** Solve $H_{\mathcal{S}^{(t)}} q^{(t)} = -[z^*]_{\mathcal{S}^{(t)}}$; remove every biactive $i \in \mathcal{S}^{(t)} \cap \mathcal{B}$ with $\sigma_i^{(t)} q_i^{(t)} \leq 0$.
3. **Stop** when no removal; output $\mathcal{S}$ and $\left(D_{\text{supp}}\right)_{ii} := \sigma_i$.

Null oracle ($\sigma \equiv 0$ on $\mathcal{B}$) coincides with the **Sparse-HO** selection: yields $\partial_{x_i} \Phi = 0$ at every biactive coord $\Longrightarrow$ outer iteration cannot move x_i **gradient starvation**. SC keeps the sign-coherent biactive coordinates and **unsticks hidden features**.

Bertrand et al., *Implicit differentiation for fast hyperparameter selection in non-smooth convex learning*, JMLR 23(149), 2022.

Algorithm 1 — support-reduced subgradient

Inputs. $\bar{x} \in \mathbb{R}^p$, step $\gamma \in (0, \Lambda_F^{-1})$, selection policy Π (e.g. SC).

1. **Solve lower level.** $\bar{y} \approx S(\bar{x})$; set $(\bar{u}, \bar{v}) := (\gamma\Psi(\bar{x}), \bar{y} - \gamma\nabla F(\bar{y}))$; partition into $\mathcal{J}^{\pm}, \mathcal{A}, \mathcal{B}^{\pm}$.
2. **Apply oracle.** Get $\nabla_x L$, $z^* := \nabla_y L$; run Π to obtain masks $M_{\mathcal{B}^{\pm}}$; working set $\mathcal{S} := \mathcal{J}^+ \cup \mathcal{J}^- \cup \operatorname{supp}(M_{\mathcal{B}^+}) \cup \operatorname{supp}(M_{\mathcal{B}^-})$ with sign vector σ.
3. **Reduced adjoint solve.** Form $H_{\mathcal{S}} := \gamma[\nabla^2 F(\bar{y})]_{\mathcal{S},\mathcal{S}}$ and solve $H_{\mathcal{S}} q_{\mathcal{S}} = -[z^*]_{\mathcal{S}}$ at cost $\mathcal{O}(|\mathcal{S}|^3)$ instead of $\mathcal{O}(p^3)$ — with $|\mathcal{S}| \ll p$ in well-regularized sparse models.
4. **Reconstruct.** $[g_{\text{imp}}]_{\mathcal{S}} := \gamma[J\Psi(\bar{x})]_{\mathcal{S},\mathcal{S}}(\sigma \odot q_{\mathcal{S}})$; return $h := \nabla_x L(\bar{x}, \bar{y}) + g_{\text{imp}}$.

h is the residual-enlarged subgradient that NBA / NTRBA consume.

NBA-$w\ell_1$ and NTRBA-$w\ell_1$

NBA-$w\ell_1$ — normalized subgradient

1. solve lower level $\to y_k := S(x_k)$
2. pick $\eta_k \in \partial\Phi(x_k)$ via oracle
3. $x_{k+1} := \mathrm{proj}_X(x_k - \alpha_k \eta_k / \|\eta_k\|)$

NTRBA-$w\ell_1$ — trust region

1. build local model m_k on radius Δ_k
2. $s_k := \arg\min_{\|s\| \le \Delta_k} m_{k(s)}$
3. accept/expand/contract Δ_k from ρ_k

Both methods accept **either** oracle. We report results with the **self-consistent oracle + NTRBA** combination.

5. Numerical Experiments

Calibrated degenerate dataset

Design matrix columns partitioned into **four feature groups**:

$$A = \Big[\ \underbrace{A_{\text{easy}}}_{\text{signal, easy}} \ \Big|\ \underbrace{A_{\text{dist}}}_{\text{nuisance}} \ \Big|\ \underbrace{A_{\text{hid}}}_{\text{signal, hidden}} \ \Big|\ \underbrace{A_{\text{noise}}}_{\text{background}}\ \Big]$$

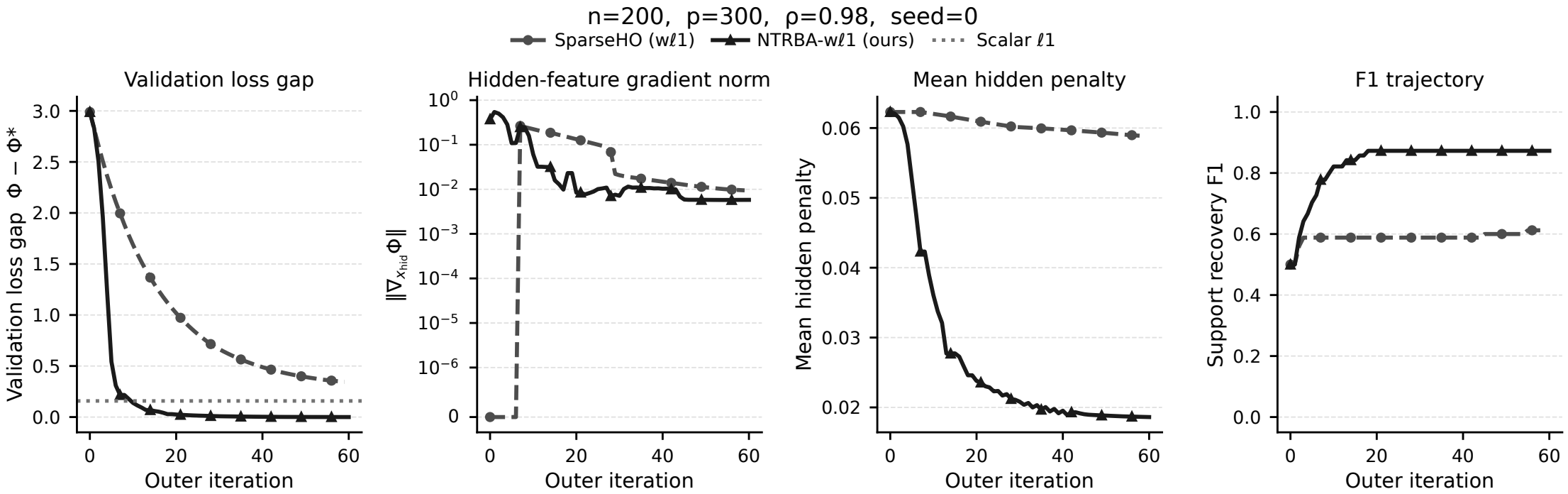
Hidden columns are **correlated with distractors**:

$$A_{\text{hid},j} = \rho\, A_{\text{dist},\pi(j)} + \sqrt{1-\rho^2}\,\xi_j, \quad \rho \in \{0.9, 0.95, 0.98\}.$$

Truth β^* supported on **easy** $\cup$ **hidden**; response $b = A\beta^* + \sigma\varepsilon$; 60/20/20 train/val/test split.

Calibration trigger. The initial hidden penalty $x^0_{\text{hid},j} = |g_{\text{hid},j}|(1+\delta)$, $\delta = 0.05$, places hidden coordinates **just inside the biactive band** — exactly where the standard implicit-diff subgradient is zero.

Gradient starvation, one instance



Gradient starvation, across configurations

Across **9 degenerate configurations**:

$\rho \in \{0.9, 0.95, 0.98\}$, $(n, p) \in \{(100, 150), (200, 300), (500, 750)\}$, mean over 5 seeds:

Method	$\|\partial_{x_{\text{hid}}} \Phi\|_{k=0} \uparrow$	Best val. loss $\downarrow$	Support $F_1 \uparrow$
Scalar ℓ_1 (grid)	— (not defined)	0.13 – 1.31	0.44 – 0.51
Sparse-HO ($w\ell_1$)	**0.00** in every config	0.14 – 5.11	0.63 – 0.70
NTRBA-$w\ell_1$ (ours)	**0.30 – 5.12**	**0.02 – 1.46**	**0.77 – 0.90**

The standard implicit-diff subgradient is **identically zero on every hidden coordinate at initialization**. The self-consistent oracle restores descent and lifts support F_1 by $\approx$ **20 pp** uniformly across scales and correlations.

Real-world classification

NTRBA **matches or improves** Sparse-HO in F_1 with **comparable sparsity**

Dataset	Method	$F_1 \uparrow$	Active % $\downarrow$	t/iter (s) $\downarrow$
MNIST 0/1	Scalar ℓ_1	**0.998**	34.5	202.4
	Sparse-HO	0.994	**1.24**	**1.9**
	NTRBA (ours)	0.994	2.99	8.2
News20	Scalar ℓ_1	**0.957**	0.500	808.3
	Sparse-HO	0.821	0.013	**4.1**
	NTRBA (ours)	0.830	**0.012**	15.2
Phishing	Scalar ℓ_1	**0.947**	81.4	50.4
	Sparse-HO	0.933	**14.7**	0.23
	NTRBA (ours)	0.935	**14.7**	**0.18**
RCV1	Scalar ℓ_1	**0.969**	6.4	254.2
	Sparse-HO	0.917	0.230	**0.10**
	NTRBA (ours)	0.920	**0.227**	0.84
Real-sim	Scalar ℓ_1	**0.951**	32.8	2335.4
	Sparse-HO	0.787	0.286	**1.03**
	NTRBA (ours)	0.790	**0.283**	2.15

Take-aways

- **Bilevel framework** for tuning per-feature weights of a weighted ℓ_1 regularizer.
- **FB reformulation** of the lower level preserves the original solution set **exactly** — unlike smoothing, which biases the upper level.
- **Computable limiting subgradient** via the coderivative of the residual map, with kernel $\{0\}$ everywhere.
- A **self-consistent biactive oracle** that resolves **gradient starvation** — the direct cause of weak hidden-feature recovery in standard implicit-diff baselines.

Thank you

Thank you for your attention!

https://david.villacis.net