

A Variational Analysis Approach for Bilevel Hyperparameter Optimization with Sparse Regularization

David Villacís



joint work with Pedro Pérez-Aros and Emilio Vilches

Bilevel Optimization Society Webinar — September 1, 2026

1. Learning Thousands of Hyperparameters

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

$$\min_{y \in Y} F(y) + G_x(y), \quad G_x(y) = \sum_{i=1}^p \psi(x_i) |y_i|, \quad \psi(x_i) = \exp(x_i).$$

Why one weight **per feature** instead of a single scalar? **Robustness** to structured disturbances, **adaptivity** — each feature gets exactly the penalty it deserves — and graceful degradation under noise.

The same weighted- ℓ_1 template covers:

- **Overparametrized regression** — $F(y) = \frac{1}{2} \| A^{\text{train}} y - b^{\text{train}} \|^2$, with $n \ll p$;
- **Wavelet image restoration** — $F(y) = \frac{1}{2} \| RWy - b^{\text{damaged}} \|^2$, R blur, W wavelet basis;
- **Sparse classification** — F = binary cross-entropy.

The hyperparameter is now a **vector**: $x \in \mathbb{R}^p$, with $p \sim 10^3 - 10^5$.

Grid search and Bayesian optimization break down at this dimension. Make x the decision variable of an **outer** problem, scored on the **validation set**:

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

If S were **single-valued and C^1** , this collapses to $\min_{x \in X} \Phi(x) := L(x, S(x))$ with the IFT hypergradient

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

But G_x is not differentiable. S need not be either — and smoothing $|\cdot|$ shifts the minimizer **off** the ℓ_1 boundary, **biasing** the upper level.

Sparse-HO (Bertrand et al., 2022): proximal **implicit differentiation** restricted to the support. Fast and accurate — but its analysis assumes **strict complementarity**.

At a **biactive** coordinate ($y_i = 0$, yet sitting **exactly on** the soft-thresholding threshold) it returns a **zero** subgradient: the outer loop **can never move that weight**. We call this **gradient starvation**.

This talk. Variational analysis keeps the subgradient **informative** exactly where Sparse-HO goes blind — and **reduces to Sparse-HO** everywhere else.

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

2. An Exact Reformulation

Introduce the **forward-backward** operator and its **residual**:

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

where $\mathcal{T}$ is **component-wise soft-thresholding** and $\mathcal{H}^\gamma$ collects the smooth data:

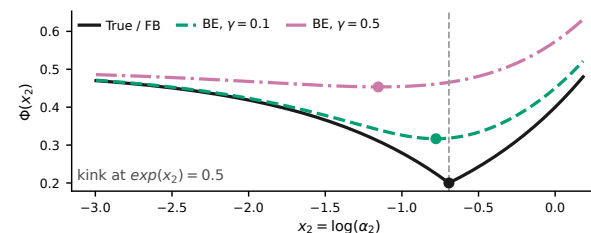
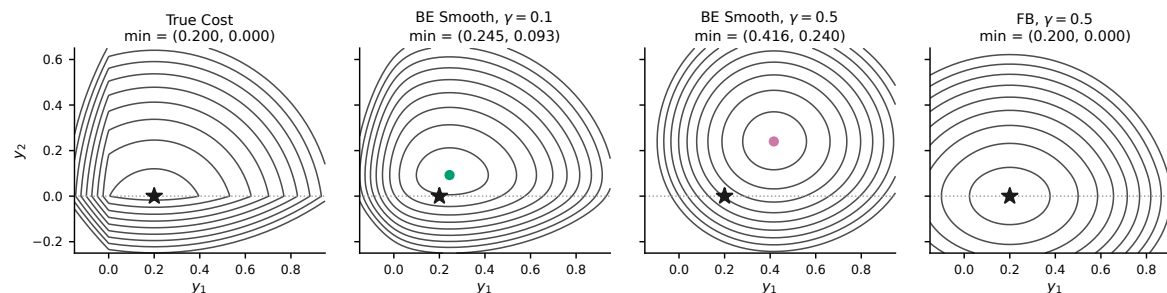
$$\mathcal{T}(u, v) := (\text{sign}(v_i) \max(|v_i| - u_i, 0))_{i=1}^p, \quad \mathcal{H}^\gamma(x, y) := (\gamma \Psi(x), y - \gamma \nabla F(y)).$$

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

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

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

Theorem: (Theorem 1.) Under strong convexity of F , $S(x)$ is **single-valued** for every $x \in X$, and $S : X \rightarrow Y$ is **locally Lipschitz**.



Smoothing $|\cdot|$ (Berkovich–Engelman $\sqrt{y_i^2 + \gamma^2}$) **displaces** the minimizer; the FB reformulation does not.

$\Rightarrow$ a genuine single-level problem
 $\Phi(x) = L(x, S(x))$, with the lower-level
 solution set left **exactly** as it was — for
every $\gamma > 0$.

3. Where the Subgradient Hides

Lemma: Let L be strictly differentiable at $(\bar{x}, \bar{y})$, $\bar{y} := S(\bar{x})$. Then

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

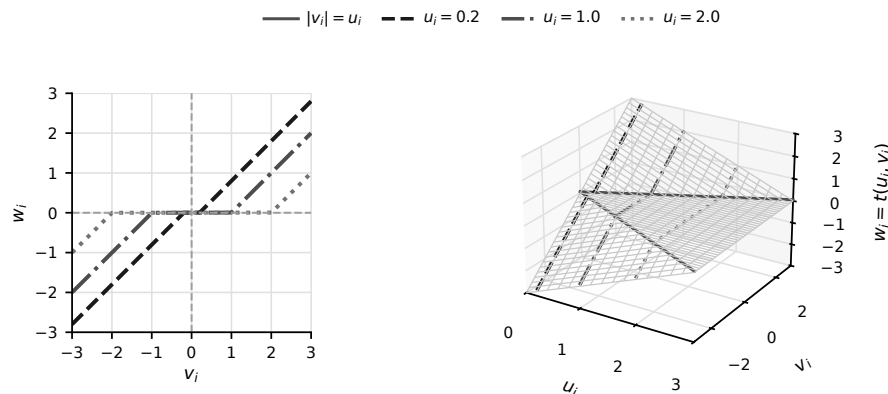
D^*S — the **Mordukhovich coderivative** — lives on $\text{gph } S$, which we cannot touch directly. But $S(x) = \{y \mid R^\gamma(x, y) = 0\}$, so we pull everything back through the **residual**:

$$\partial_R^\gamma \Phi(\bar{x}) := \nabla_x L + \{x^* \mid (x^*, -\nabla_y L) \in D^*R^\gamma(\bar{x}, \bar{y})(z^*), \exists z^*\} \supseteq \partial\Phi(\bar{x}).$$

The **residual-enlarged subdifferential**: a tractable outer approximation, **exact** when R^γ is lower-regular.

Plan: differentiate the residual R^γ — every piece of it is explicit.

The graph of $\mathcal{T}$ is **piecewise linear**, so the limiting normal cone of $\text{gph } \mathcal{T}$ — and hence D^*R^γ — is available in **closed form**.



On the smooth pieces the normal cone is a **line**: $D^*\mathcal{T}$ is **single-valued**, no choice to make.

At a **biactive** kink ($\bar{v}_i = \pm \bar{u}_i$) it is a **union of three** branches — on $\mathcal{B}^+$,

- $(u_i^*, v_i^*) = (0, 0)$ — **drop** i ;
- $(u_i^*, v_i^*) = (-t, t)$, $t \in [0, z_i^*]$ — **interpolate**;
- $(u_i^*, v_i^*) = (-z_i^*, z_i^*)$ — **activate** i .

The selector $D_{\text{supp}} \in \{-1, 0, +1\}^{p \times p}$ is **forced off the biactive set, free on it**. That freedom is the oracle.

The coderivative sum/chain rules split the residual through $\nabla \mathcal{H}^\gamma$ and $D^*\mathcal{T}$:

$$D^*R^\gamma(x, y)(z^*) = (0, z^*) + \nabla \mathcal{H}^\gamma(x, y)^\top \circ D^*\mathcal{T}(\mathcal{H}^\gamma(x, y))(-z^*).$$

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

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

And the qualification condition costs nothing here: $\ker D^*R^\gamma(x, y) = \{0\}$ holds for **every** (x, y) and $\gamma \in (0, \Lambda_F^{-1})$.

A hyper-subgradient = a sign choice per biactive coordinate + one SPD linear solve.

4. The Sign-Consistent Oracle

Partition the coordinates by their FB complementarity status at $(\bar{u}, \bar{v}) = \mathcal{H}^\gamma(\bar{x}, \bar{y})$:

$$\underbrace{\mathcal{J}^\pm}_{\text{support } (\bar{y}_i \neq 0)} \quad \underbrace{\mathcal{A}}_{\text{strictly inactive}} \quad \underbrace{\mathcal{B}^\pm}_{\text{biactive } (\bar{v}_i = \pm \bar{u}_i)}$$

- On $\mathcal{J}^\pm, \mathcal{A}$ the coderivative $D^*\mathcal{T}$ is **single-valued** — no choice to make.
- On $\mathcal{B}^\pm$ it is genuinely **set-valued** — the **union of three** branches.

Support-restricted implicit differentiation (the null oracle $\equiv$ Sparse-HO) assigns $\partial_{x_i} \Phi = \mathbf{0}$ to every biactive coordinate $\implies$ the outer loop **cannot move** it. Structurally so at $\alpha_{\max}$, where the support is empty and **every** coordinate is biactive.

Write $\sigma_i := (D_{\text{supp}})_{ii} \in \{-1, 0, +1\}$ for the sign the oracle assigns to coordinate i , and let q solve the reduced adjoint system. **Keep** a biactive i only if $\sigma_i q_i > 0$ — 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 $(D_{\text{supp}})_{ii} := \sigma_i$.

Terminates in $\leq |\mathcal{B}|$ steps. When $\nabla^2 F(\bar{y})$ is **diagonal**, the initialization is already sign-consistent — no pruning.

Proposition: (Descent test.) For a biactive coordinate i , let h be the SC subgradient with reduced-adjoint solution q . Then $\Phi'(\bar{x}; -e_i) = -h_i$; hence $-e_i$ is a **strict descent direction** for Φ **if and only if** $\sigma_i q_i > 0$.

Proposition: (Reduction.) Under **strict complementarity** ($\mathcal{B}^+ = \mathcal{B}^- = \emptyset$) the selector is forced, $(D_{\text{supp}})_{ii} = \text{sign}(\bar{y}_i)$, the adjoint decouples on the support, and $\partial_R^\gamma \Phi(\bar{x}) = \{\nabla \Phi_{\text{Sparse-HO}(\bar{x})}\}$.

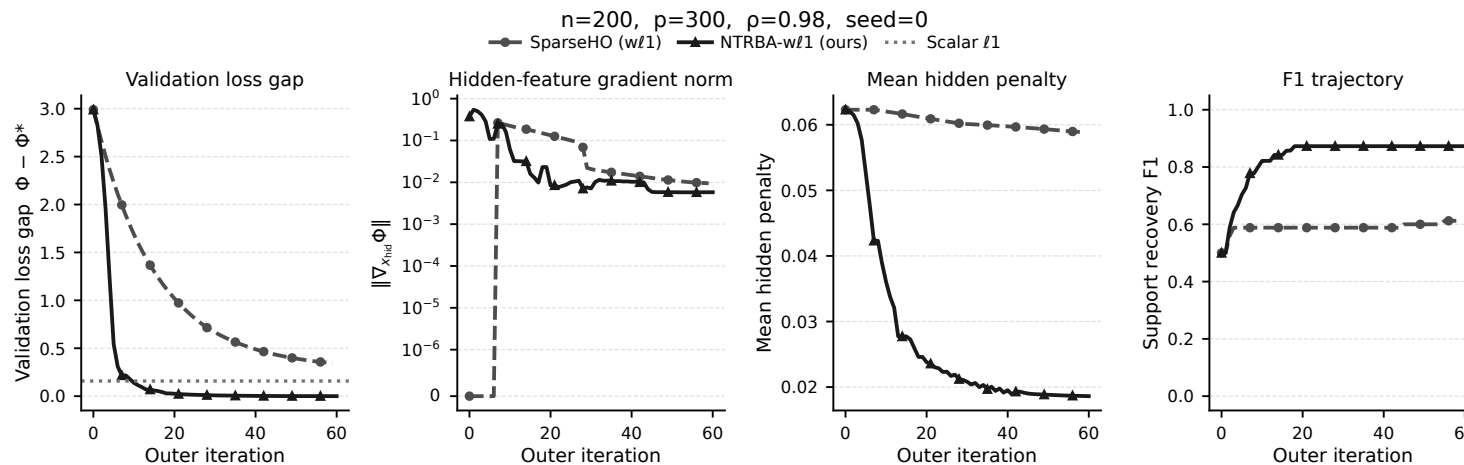
Algorithm 1. Partition $\mathcal{I}^\pm, \mathcal{A}, \mathcal{B}^\pm \rightarrow$ oracle fixes σ and the working set $\mathcal{S} \rightarrow$ one reduced solve $H_{\mathcal{S}} q_{\mathcal{S}} = -[\nabla_y L]_{\mathcal{S}}$ at $\mathcal{O}(|\mathcal{S}|^3)$, $|\mathcal{S}| \ll p \rightarrow$ reconstruct $h := \nabla_x L + \gamma[J\Psi(\bar{x})]_{\mathcal{S}, \mathcal{S}}(\sigma \odot q_{\mathcal{S}})$. Drives either outer loop: normalized-subgradient (NBA) or trust-region (NTRBA).

Identical to Sparse-HO off the degenerate set, corrective on $\mathcal{B}^\pm$.

5. Does It Work?

Four feature groups, with the hidden signal columns **correlated with the distractors**:

$$A = \left[\begin{array}{c|c|c|c} \underbrace{A_{\text{easy}}}_{\text{signal, easy}} & \underbrace{A_{\text{dist}}}_{\text{nuisance}} & \underbrace{A_{\text{hid}}}_{\text{signal, hidden}} & \underbrace{A_{\text{noise}}}_{\text{background}} \end{array} \right], \quad A_{\text{hid},j} = \rho A_{\text{dist},\pi(j)} + \sqrt{1 - \rho^2} \xi_j.$$



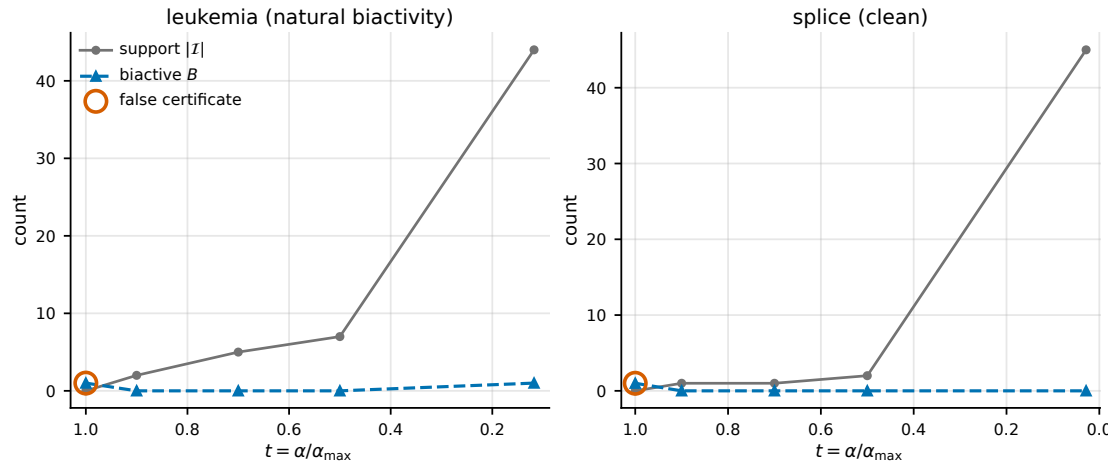
Calibration trigger. The initial hidden penalty $x_{\text{hid},j}^0 = |g_{\text{hid},j}|(1 + \delta)$, $\delta = 0.05$, places hidden coordinates **just inside the biactive band** — exactly where the standard implicit-diff subgradient is **zero**.

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 sign-consistent oracle restores descent and lifts support F_1 by ≈ 20 pp uniformly across scales and correlations. At real-world scale (MNIST, News20, RCV1, Real-sim, Phishing) NTRBA **matches** Sparse-HO in F_1 and sparsity.

At the **regularization-path anchor** $\alpha_{\max}$, the support is **empty** and the top-gradient feature is **biactive by definition** — the warm start of every path solver.



Across **9 benchmark datasets**, the null oracle returns $\|g_{\text{null}}\| = \mathbf{0}$ — a **false** stationarity certificate — yet the true slope $\Phi'(\bar{x}; -e_{i^*}) \in [-0.36, -0.03] < 0$. The SC estimate matches it to within $\approx \mathbf{1.5\%}$.

6. Conclusions

- **Exact, not smoothed.** The FB reformulation keeps the lower-level solution set unchanged — no upper-level bias.
- **Computable, not abstract.** The coderivative of soft-thresholding is closed-form; a hyper-subgradient costs one reduced SPD solve, $\mathcal{O}(|\mathcal{S}|^3) \ll \mathcal{O}(p^3)$.
- **Principled, not heuristic.** The sign-consistent oracle keeps a biactive coordinate **iff** $-e_i$ is a true descent direction — and it **reduces exactly to Sparse-HO** under strict complementarity.
- **It matters in practice.** Restores descent where Sparse-HO starves ($F_1 + 20$ pp), matches it at real-world scale, and replaces a **false** stationarity certificate at $\alpha_{\max}$ with a **sound** one.

Thank you — questions welcome!



<https://david.villacis.net>