LEARNING PIECEWISE-SMOOTH IMAGE STRUCTURE WITH IMPLICIT REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In image regression, coefficient images are often piecewise smooth: approximately constant or smoothly varying within regions, with abrupt changes across boundaries. Explicit penalties such as total variation (TV) exploit this structure for spatially coherent, interpretable estimates but require a nonsmooth fit for every tuning value. To learn it through implicit regularization instead, we decompose the coefficient image into a polynomial background and a component reconstructed from local differences of a chosen order. With the background profiled out, a difference-of-squares factorization encourages sparse differences and a quadratic term keeps them consistent with an image. Gradient descent from a small balanced initialization, with backtracking and validation-based early stopping, fits the factors. Under correlated sub-Gaussian designs, we establish finite-step prediction and coefficient-error bounds. Under bounded variation, conservative steps yield an $n^{-1/2}$ prediction rate up to logarithmic and geometric factors, without restricted-eigenvalue assumptions. A validation oracle inequality supports checkpoint and order selection at logarithmic cost in the candidate count. The small-initialization gradient-flow limit connects the method to minimum-variation interpolation. Simulations and a sea surface temperature application show competitive accuracy at lower computational cost than TV in the studied settings.

1 INTRODUCTION

Image regression relates an image-valued predictor to a scalar response through a coefficient image and has been developed for applications in neuroimaging, medical imaging, and environmental science (Zhou et al., 2013; Wang & Zhu, 2017; Kang et al., 2018; Feng et al., 2021; Wu & Feng, 2023; Liao et al., 2026). A central feature of many coefficient images is *piecewise-smooth structure*: the coefficient varies smoothly, or is well approximated by a low-degree polynomial, within spatial regions but may change abruptly across their boundaries. This structure is natural when neighboring voxels represent the same anatomical, functional, or physical region. By borrowing information among neighboring voxels while preserving localized changes, piecewise-smooth modeling can improve estimation and produce spatially coherent, interpretable coefficient images for identifying response-associated regions (Wang & Zhu, 2017; Feng et al., 2021).

Existing methods for capturing this structure broadly follow two strategies. The first represents the coefficient image through a lower-dimensional model or a structured stochastic prior, including low-rank decompositions, spatial variable-selection priors, and thresholded Gaussian processes (Zhou et al., 2013; Goldsmith et al., 2014; Kang et al., 2018; Lin et al., 2026). These methods reduce the effective dimension or encourage spatially coherent estimates, but they rely on a selected rank, prior covariance, or global representation and may not adapt directly to irregular local boundaries (Zhou et al., 2013; Goldsmith et al., 2014; Kang et al., 2018). The second strategy penalizes local differences: first-order total variation favors piecewise-constant images, while higher-order differences accommodate piecewise-linear and piecewise-polynomial regions (Rudin et al., 1992; Kim et al., 2009; Tibshirani, 2014; Sadhanala et al., 2017; Wang & Zhu, 2017; Feng et al., 2021). Variation penalties smooth within regions while preserving edges, but their computation becomes demanding for large three- or higher-dimensional images because each tuning value requires solving a nonsmooth problem involving the full image (Wang & Zhu, 2017; Feng et al., 2021). These observations motivate our main question: *can a simple procedure capture piecewise-smooth coefficient images without repeatedly solving an explicitly variation-penalized problem?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We answer this question by retaining the geometry of local differences while moving the sparsity-inducing mechanism from the objective to the parameterization. Let $\boldsymbol{b} = \mathrm{vec}(\mathcal{B})$ denote the vectorized coefficient image and let D_k collect its $(k+1)$-st local grid differences; we call k the order index. We decompose $\boldsymbol{b}$ as $\boldsymbol{b} = Q_k\boldsymbol{a} + D_k^\dagger \boldsymbol{w}$, where $Q_k\boldsymbol{a}$ is a polynomial background in the null space of D_k, and $D_k^\dagger \boldsymbol{w}$ reconstructs the component described by local differences. Varying k allows the same decomposition to represent piecewise-constant, piecewise-linear, and piecewise-polynomial coefficient images (Tibshirani, 2014; Sadhanala et al., 2017). For each $\boldsymbol{w}$, the background coefficient $\boldsymbol{a}$ has a closed-form least-squares solution. A free $\boldsymbol{w}$ need not equal the local differences of $D_k^\dagger \boldsymbol{w}$, so we add the quadratic consistency term $\gamma\|(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w}\|_2^2$ (cf. Pérez et al., 2003), which vanishes exactly when $\boldsymbol{w} = D_k\boldsymbol{b}$ for some $\boldsymbol{b} \in \mathbb{R}^p$. Its weight γ is set from the training design without validation tuning. To encourage sparse local differences without an explicit ℓ_1 penalty, we write $\boldsymbol{w} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$ and run GD from the balanced initialization $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha\mathbf{1}$. Similar parameterizations induce implicit regularization in high-dimensional models (Vaškevičius et al., 2019; Woodworth et al., 2020; Zhao et al., 2022; Li et al., 2023; Sui et al., 2023); here the induced sparsity acts on local image differences. Backtracking selects a stable step size, and an independent validation set chooses a stopping point from checkpoints fixed before training. Thus, one GD path provides all candidates instead of repeatedly solving a nonsmooth TV problem over a penalty grid.

Our theoretical analysis keeps finite-step GD distinct from gradient flow, whose implicit biases can differ (Nacson et al., 2022; Even et al., 2023). Under correlated sub-Gaussian predictor images and sub-Gaussian noise, we first derive prediction and coefficient-error bounds for the discrete GD iterates. The prediction bound depends on the variation of the coefficient error and an explicit remainder measuring the effect of discrete factor updates. Its structure parallels nonasymptotic bounds for explicitly variation-regularized image regression: smaller local variation yields a sharper bound even though our training objective contains no variation penalty (Wang & Zhu, 2017; Liao et al., 2026). Under bounded variation of the true coefficient image and a conservative step-size rule, we further show that every iterate in an explicit iteration window attains an $n^{-1/2}$ prediction rate up to logarithmic and geometric factors. This requires neither the restricted-isometry condition of Zhao et al. (2022) nor the fixed-dimensional regime of Zhou et al. (2013), and it allows $p > n$ (Appendix B.2). Predetermined checkpoints and independent validation preserve the rate with only logarithmic dependence on the number of candidates and permit selection among several orders. Finally, we analyze gradient flow separately and show that, after convergence and as initialization vanishes, its limit points are minimum-$\|D_k\boldsymbol{b}\|_1$ interpolating solutions. This result explains the connection to variation penalties without equating practical finite-step GD with an explicitly penalized estimator.

We evaluate the method through three-dimensional simulations and a real-data application. The main experiments compare GD with TV regression, low-rank CP regression, ridge regression, and elastic net across several piecewise-constant coefficient images. GD matches or improves on TV in prediction and coefficient recovery while requiring substantially less fit-and-selection time on the larger grids. Additional experiments vary the consistency weight γ, initialization α, and order index k, and examine the complete GD paths. The results show stable performance across positive consistency weights and sufficiently small initializations, while the preferred order depends on both within-region smoothness and boundary jumps. For sea surface temperature (SST) data (Liao et al., 2026), GD and TV yield the lowest test MSE, with GD taking less than half of TV's runtime. Appendix A provides supplementary experiments, Appendix B gives further theoretical analysis, and Appendix C contains proofs of the main results.

2 Problem Setup and Piecewise-Smooth Image Structure

We first specify the image regression model and the local differences used to describe piecewise-smooth structure. In this paper, multidimensional image arrays are denoted by $\mathcal{X}_i$ and $\mathcal{B}$, and matrices by uppercase letters such as X. Vectors use bold italic lowercase letters, such as $\boldsymbol{x}_i$, whereas their entries and other scalars are unbolded, such as x_{ij} and y_i. The operators vec and † denote vectorization and the Moore–Penrose pseudoinverse; I_j is the $j \times j$ identity matrix. We use $\|\cdot\|_q$ for vector ℓ_q norms, $\|\cdot\|_F$ for the Frobenius norm, and $\|\cdot\|_{\mathrm{op}}$ for the matrix spectral norm; $\langle\cdot,\cdot\rangle$ denotes Euclidean or Frobenius inner products. Powers and products marked by $\odot$ are coordinatewise.

Suppose we observe $\{(\mathcal{X}_i, y_i)\}_{i=1}^n$, where $\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is an image-valued predictor and $y_i \in \mathbb{R}$ is a scalar response. We assume a scalar-on-image regression model (Zhou et al., 2013; Wang & Zhu, 2017; Bai et al., 2027):

$$y_i = \langle \mathcal{X}_i, \mathcal{B}^\star \rangle + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\mathcal{B}^\star \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the coefficient image on a d-dimensional grid and ε_i is the noise. Set $p = \prod_j p_j$, $\boldsymbol{x}_i = \mathrm{vec}(\mathcal{X}_i)$, $\boldsymbol{b}^\star = \mathrm{vec}(\mathcal{B}^\star)$, and $\boldsymbol{b} = \mathrm{vec}(\mathcal{B})$, using the same voxel ordering. Taking $\boldsymbol{x}_i^\top$ as row i of X gives $\boldsymbol{y} = X\boldsymbol{b}^\star + \boldsymbol{\varepsilon}$. For clarity, we focus on linear image regression; the parameterization below also accommodates generalized linear model losses.

Fix an order index $k \geq 0$ with all $p_j \geq k+2$. Stack the valid, unscaled $(k+1)$-st forward differences:

$$D_k = \begin{bmatrix} \Delta_1^{k+1} \\ \vdots \\ \Delta_d^{k+1} \end{bmatrix} \in \mathbb{R}^{m_k \times p}, \qquad m_k = \sum_j (p_j - k - 1) \prod_{\ell \neq j} p_\ell. \tag{1}$$

The jth block acts as $(\Delta_j^{k+1}\boldsymbol{b})_{\boldsymbol{i}} = \sum_{r=0}^{k+1} (-1)^{k+1-r} \binom{k+1}{r} b_{\boldsymbol{i}+r\boldsymbol{e}_j}$ for $1 \leq i_j \leq p_j - k - 1$, with the other voxel coordinates in their full ranges, where $b_{\boldsymbol{i}}$ is the image value at voxel $\boldsymbol{i} = (i_1, \ldots, i_d)$ and $\boldsymbol{e}_j$ is one grid step along axis j. Different orders k (Sadhanala et al., 2017) capture different local structures in the coefficient image $\mathcal{B}$, as listed below and illustrated for $k = 0, 1, 2$ in Figure 1:

- $k = 0$ (piecewise constant): $\|D_0\boldsymbol{b}\|_1 = \sum_{j,\boldsymbol{i}} |b_{\boldsymbol{i}+\boldsymbol{e}_j} - b_{\boldsymbol{i}}|$ is anisotropic TV (Wang & Zhu, 2017) and vanishes on constant regions.
- $k = 1$ (piecewise linear): $\|D_1\boldsymbol{b}\|_1 = \sum_{j,\boldsymbol{i}} |b_{\boldsymbol{i}+2\boldsymbol{e}_j} - 2b_{\boldsymbol{i}+\boldsymbol{e}_j} + b_{\boldsymbol{i}}|$ is the ℓ_1 trend-filtering penalty (Kim et al., 2009) and vanishes on coordinatewise linear regions.
- $k \geq 2$ (piecewise polynomial structure): for $k = 2$, $\|D_2\boldsymbol{b}\|_1 = \sum_{j,\boldsymbol{i}} |b_{\boldsymbol{i}+3\boldsymbol{e}_j} - 3b_{\boldsymbol{i}+2\boldsymbol{e}_j} + 3b_{\boldsymbol{i}+\boldsymbol{e}_j} - b_{\boldsymbol{i}}|$ vanishes on coordinatewise quadratic regions; higher orders follow analogously, as in one-dimensional higher-order trend filtering (Tibshirani, 2014).

3 Gradient Descent with Difference Reparameterization

In this section, we introduce a GD algorithm that uses local differences to capture piecewise-smooth structure in the coefficient image. We decompose the coefficient image into a polynomial background and a component reconstructed from its local differences, then fit the difference coordinates through a difference-of-squares factorization.

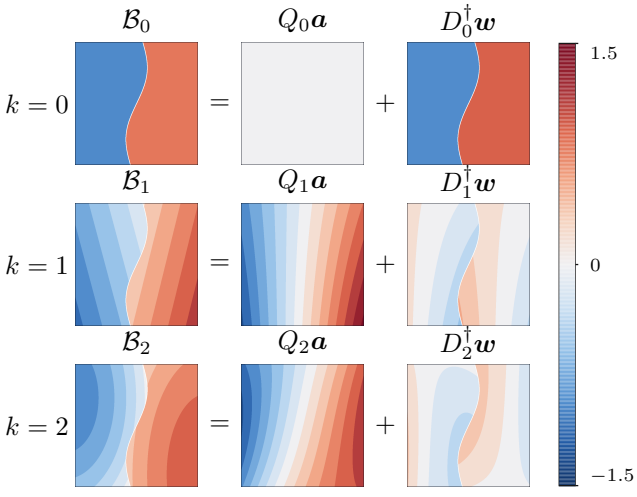


Figure 1: Piecewise-smooth coefficient images and their decompositions for $k = 0, 1, 2$.

Coefficient image decomposition. The difference operator leaves a low-degree polynomial component undetermined, as in the generalized-lasso treatment of difference penalties (Tibshirani & Taylor, 2011). Let $Q_k \in \mathbb{R}^{p \times r_k}$ have orthonormal columns spanning $\ker D_k$, where $r_k = (k+1)^d$. For any coefficient vector $\boldsymbol{b}$, define its background coordinates $\boldsymbol{a} = Q_k^\top \boldsymbol{b}$ and local differences $\boldsymbol{w} = D_k\boldsymbol{b}$. Since $D_k^\dagger D_k = I_p - Q_k Q_k^\top$, every coefficient vector satisfies

$$\boldsymbol{b} = Q_k\boldsymbol{a} + D_k^\dagger \boldsymbol{w}. \tag{2}$$

In this decomposition (Figure 1, right), $Q_k\boldsymbol{a}$ is the unpenalized polynomial background, while $\boldsymbol{w}$ records local differences and $D_k^\dagger\boldsymbol{w}$ reconstructs their image component. In

three dimensions, $r_k = 1, 8, 27$ for $k = 0, 1, 2$, and the $k = 1$ background includes xyz as well as linear terms. If $\boldsymbol{a}$ and $\boldsymbol{w}$ are treated as free coordinates, the reconstruction still makes sense, but its differences equal $\boldsymbol{w}$ precisely when $(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w} = \mathbf{0}$, because $D_k D_k^\dagger$ projects onto $\mathrm{range}(D_k)$. Appendix C.1 proves these identities.

With the decomposition in (2), we minimize the training objective $\frac{1}{2n}\|X(Q_k\boldsymbol{a} + D_k^\dagger\boldsymbol{w}) - \boldsymbol{y}\|_2^2 + \frac{\gamma}{2}\|(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w}\|_2^2$ over free $\boldsymbol{a}$ and $\boldsymbol{w}$, with $\gamma \geq 0$. The first term is the empirical prediction loss; the second is a quadratic consistency term (cf. Pérez et al., 2003). It vanishes at the differences of the true coefficient image, $\boldsymbol{w}^\star = D_k\boldsymbol{b}^\star$.

For each fixed $\boldsymbol{w}$, the minimum-norm least-squares polynomial background coefficient is $\boldsymbol{a} = (XQ_k)^\dagger(\boldsymbol{y} - XD_k^\dagger\boldsymbol{w})$; Appendix C.1.1 gives the derivation. Let $\Pi_k = I_n - (XQ_k)(XQ_k)^\dagger$ be the residual projection, and set $\widetilde{X}_k = \Pi_k X D_k^\dagger/\sqrt{n}$ and $\widetilde{\boldsymbol{y}}_k = \Pi_k\boldsymbol{y}/\sqrt{n}$. The reconstructed coefficient vector is $\boldsymbol{b}(\boldsymbol{w}) = Q_k\boldsymbol{a} + D_k^\dagger\boldsymbol{w}$, and the profiled loss is

$$\mathcal{L}_k(\boldsymbol{w}) = \tfrac{1}{2}\|\widetilde{X}_k\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k\|_2^2 + \tfrac{\gamma}{2}\|(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w}\|_2^2.$$

Profiling applies the fixed projection Π_k, with no support-dependent refitting. The algorithm requires $\mathrm{rank}(XQ_k) = r_k < n$; a full-row-rank X retains $n - r_k$ residual directions. Appendix C.1.1 gives the exact rank identity and the rank-deficient case.

The consistency weight γ encourages $\boldsymbol{w}$ to match the differences of the reconstructed coefficient image; Appendix B.1 analyzes its effect on the loss and GD paths. In practice, we set $\gamma = \|\widetilde{X}_k\|_{\mathrm{op}}^2$, which gives the data-fit term and the consistency term the same maximal curvature and uses only the training design, with no validation-based tuning. The sensitivity analysis in Section 5 shows similar errors across positive weights for four coefficient images, but larger errors at $\gamma = 0$.

GD with difference reparameterization. We parameterize the difference coordinates as $\boldsymbol{w} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$. Similar parameterizations are employed in sparse estimation (Vaškevičius et al., 2019; Woodworth et al., 2020; Sui et al., 2023), where small initialization can produce gradual coordinate activation (Pesme & Flammarion, 2023). Although $D_k^\dagger$ discards inconsistent coordinates in reconstruction, factor updates couple them to consistent ones; hence γ changes finite-step paths. Initialize $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha\mathbf{1}_{m_k}$ with $\alpha > 0$, giving $\boldsymbol{w}_0 = \mathbf{0}$. The loss gradient is $\nabla\mathcal{L}_k(\boldsymbol{w}) = \widetilde{X}_k^\top(\widetilde{X}_k\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k) + \gamma(\boldsymbol{w} - D_k D_k^\dagger\boldsymbol{w})$. Let $\boldsymbol{g}_t = \nabla\mathcal{L}_k(\boldsymbol{w}_t)$. Factor GD gives

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t \odot (\mathbf{1}_{m_k} - 2\eta_t\boldsymbol{g}_t), \qquad \boldsymbol{v}_{t+1} = \boldsymbol{v}_t \odot (\mathbf{1}_{m_k} + 2\eta_t\boldsymbol{g}_t). \tag{3}$$

We choose each step by backtracking, shrinking η_t until the factor updates preserve positivity and sufficiently decrease the loss (Armijo, 1966):

$$2\eta_t\|\boldsymbol{g}_t\|_\infty < 1, \qquad \mathcal{L}_k(\boldsymbol{w}_{t+1}) \leq \mathcal{L}_k(\boldsymbol{w}_t) - 0.0004\eta_t\sum_e (u_{t,e}^2 + v_{t,e}^2)g_{t,e}^2. \tag{4}$$

The first inequality keeps both multipliers in (3) positive. In the second, $4\sum_e(u_{t,e}^2 + v_{t,e}^2)g_{t,e}^2$ is the squared norm of the factor gradient, so the required decrease scales with the proposed step and that gradient. Sufficiently small steps pass both tests. Appendix A.5 specifies the backtracking procedure, while Appendix C.3 shows how a numerical acceptance tolerance changes the finite-step bounds. We use validation error to select among fixed GD checkpoints and the background-only fit. Algorithm 1 summarizes the complete procedure.

Computational complexity. Each GD iteration costs $O(np + p\sum_j p_j + d(k+1)p + nr_k)$: np for multiplication by X and $X^\top$, $p\sum_j p_j$ for reconstructing the coefficient image through $D_k^\dagger$, $d(k+1)p$ for computing local differences, and nr_k for updating the polynomial background. For explicit TV, a standard first-order primal–dual method costs $O(np + dp)$ per iteration with a dense design and local grid differences (Chambolle & Pock, 2011). Selecting among L penalty weights therefore costs $O\{(np+dp)\sum_{\ell=1}^L I_\ell\}$, where I_ℓ is the number of iterations for weight ℓ. Our single GD trajectory supplies every early-stopping candidate, giving a computational advantage when TV requires a long multi-penalty path, especially on large three-dimensional grids. Appendix A.5 details the computational comparison between GD and TV.

Algorithm 1 GD with validation-based early stopping

Require: Training and validation data, grid, candidate orders $\mathcal{K}$, α, budget $T_{\max}$, checkpoints $\mathcal{T}$

for $k \in \mathcal{K}$ **do**
  Require $p_j \geq k+2$ for every j and $\operatorname{rank}(XQ_k) = r_k < n$
  Construct $D_k, Q_k, \widetilde{X}_k, \widetilde{\boldsymbol{y}}_k$; set $\gamma = \|\widetilde{X}_k\|_{\mathrm{op}}^2$
  Initialize $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \mathbf{1}_{m_k}$; retain the background-only fit $\boldsymbol{b}(\mathbf{0})$
  if $\gamma = 0$ **then continue** to the next order
  for $t = 0, \ldots, T_{\max} - 1$ **do**
    Compute $\boldsymbol{w}_t = \boldsymbol{u}_t^{\odot 2} - \boldsymbol{v}_t^{\odot 2}$ and $\boldsymbol{g}_t = \nabla \mathcal{L}_k(\boldsymbol{w}_t)$
    Backtrack as in Appendix A.5; on success update using (3)
    if $t+1 \in \mathcal{T}$ **then** retain $\boldsymbol{b}(\boldsymbol{w}_{t+1})$ as a candidate
  end for
end for
return the candidate with the smallest validation MSE

4 Theoretical Guarantees

In this section, we analyze the proposed procedure. We first establish finite-step prediction and coefficient-error bounds, then derive an explicit prediction rate and analyze validation-based early stopping, and finally connect the small-initialization gradient-flow limit to minimum-variation interpolation. Appendix B develops additional theoretical results, and Appendix C contains the complete proofs. For a new predictor with covariance Σ, define population prediction risk $\mathcal{R}(\boldsymbol{b}) = (\boldsymbol{b} - \boldsymbol{b}^\star)^\top \Sigma (\boldsymbol{b} - \boldsymbol{b}^\star)$. The coefficient error is $\|\mathcal{B} - \mathcal{B}^\star\|_F^2 = \|\boldsymbol{b} - \boldsymbol{b}^\star\|_2^2$. We impose the following assumptions throughout this section.

Assumption 1 (Conditional noise)**.** For a deterministic $\sigma \geq 0$ and every $\boldsymbol{t} \in \mathbb{R}^n$, $\mathbb{E}[\exp(\boldsymbol{t}^\top \boldsymbol{\varepsilon}) \mid X] \leq \exp(\sigma^2 \|\boldsymbol{t}\|_2^2/2)$.

Assumption 2 (Random design)**.** The rows $\boldsymbol{x}_i$ are independent and identically distributed, have mean zero and covariance $\Sigma \succ 0$, and satisfy $\mathbb{E}\exp(t \boldsymbol{x}_i^\top \boldsymbol{h}) \leq \exp(\omega^2 t^2 \boldsymbol{h}^\top \Sigma \boldsymbol{h}/2)$ for every $\boldsymbol{h} \in \mathbb{R}^p, t \in \mathbb{R}$, with $\omega \geq 1$.

Assumption 1 requires the full noise vector to be sub-Gaussian conditional on X. Assumption 2 requires sub-Gaussian predictors with covariance $\Sigma \succ 0$ but allows their coordinates to be correlated. Define the prediction scale $\ell_k = \max_{1 \leq j \leq m_k} \|\Sigma^{1/2} D_k^\dagger \boldsymbol{e}_j\|_2$, where $\boldsymbol{e}_j$ is the jth basis vector of $\mathbb{R}^{m_k}$. It is the largest prediction norm of the coefficient vector reconstructed from a unit change in one difference coordinate.

4.1 Finite-step prediction and estimation

For the Tth GD iterate, write $\boldsymbol{h}_T = \boldsymbol{b}(\boldsymbol{w}_T) - \boldsymbol{b}^\star$ for its coefficient error and recall that $\boldsymbol{w}^\star = D_k \boldsymbol{b}^\star$. The bound below uses $\|D_k \boldsymbol{h}_T\|_1$, which measures how much the coefficient error varies across the image, and the nonnegative term $\mathcal{E}_T(\boldsymbol{w}^\star)$, defined in Appendix C.3, equation (26), which bounds how much larger $\mathcal{L}_k(\boldsymbol{w}_T)$ can be than $\mathcal{L}_k(\boldsymbol{w}^\star)$ after T GD steps.

Theorem 1 (Finite-step prediction bound)**.** *Under Assumptions 1–2, if $n \geq C\omega^4\{r_k + \log(8m_k/\delta)\}$, then with probability at least $1 - \delta$, simultaneously for every GD iterate $T \geq 1$ generated by (3) with step sizes satisfying (4),*

$$\mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)) \lesssim \mathcal{E}_T(\boldsymbol{w}^\star) + \sigma \ell_k \|D_k \boldsymbol{h}_T\|_1 \sqrt{\frac{\log(8m_k/\delta)}{n}} + \frac{\omega^4 \ell_k^2 \|D_k \boldsymbol{h}_T\|_1^2 \log(2m_k)}{n} + \frac{\sigma^2\{r_k + \log(8/\delta)\}}{n}.$$

The hidden constant is absolute. Dividing the right side by $\lambda_{\min}(\Sigma)$ gives the corresponding coefficient-error bound.

The term linear in $\|D_k \boldsymbol{h}_T\|_1$ describes the interaction between noise and spatial variation, the quadratic term accounts for random-design deviation, and the final term is the noise cost of estimating the polynomial background. This structure mirrors TV-based image regression theory (Wang &

Zhu, 2017): both bounds improve with smaller variation, although here GD induces this behavior without an explicit penalty. Appendices C.3 and C.5 give the detailed optimization and random-design arguments (Liaw et al., 2017). Theorem 1 still depends on $\|D_k \boldsymbol{h}_T\|_1$ and $\mathcal{E}_T(\boldsymbol{w}^\star)$ evaluated after running GD. Theorem 2 goes further: under bounded variation, it proves that GD attains an explicit prediction rate within a finite iteration window, without assuming that these two iterate-dependent quantities are already small.

4.2 An explicit finite-step prediction rate

We assume that the true coefficient image has bounded variation, $\|D_k \boldsymbol{b}^\star\|_1 \le V$. For initialization size $\alpha > 0$, define $W_\alpha = V\{2 + \log_+(V/\alpha^2)\} + 2m_k\alpha^2$, where $\log_+(x) = \max\{\log x, 0\}$ and the first term is interpreted as zero when $V = 0$. Thus W_α combines the variation bound with the effect of initialization.

Theorem 2 (Explicit finite-step prediction rate)**.** *Under Assumptions 1–2, suppose $\|D_k \boldsymbol{b}^\star\|_1 \le V$ and $n \ge C\omega^4\{r_k + \log(20m_k/\delta)\}$, where $0 < \delta < 1/4$. Use $\gamma = \|\widetilde{X}_k\|_{\mathrm{op}}^2$, the conservative step-size rule in (55), and a budget $T_{\max} \ge T_+$, where (56) defines T_- and T_+. In exact arithmetic, with probability at least $1 - \delta/2$, $\mathrm{rank}(XQ_k) = r_k$ and, if $\widetilde{X}_k \ne 0$, backtracking accepts the prescribed step without reduction through iteration T_+, and every $T_- \le T \le T_+$ satisfies*

$$\mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)) \le C\left[\sigma\ell_k W_\alpha \sqrt{\frac{\log(20m_k/\delta)}{n}} + \frac{\omega^4 \ell_k^2 W_\alpha^2 \log(2m_k)}{n} + \frac{\sigma^2\{r_k + \log(20/\delta)\}}{n}\right]. \quad (5)$$

If $\widetilde{X}_k = 0$, the background-only fit satisfies the same bound. Coefficient error is bounded by the right side divided by $\lambda_{\min}(\Sigma)$.

As in Theorem 1, the three terms in (5) account for, respectively, noise interacting with the finite-step variation scale W_α, random-design deviation, and the noise cost of estimating the r_k-dimensional polynomial background. Since W_α depends only on V, the grid size, and initialization, the leading term gives an explicit $n^{-1/2}$ prediction rate up to ℓ_k and a logarithmic factor. Up to these geometric and logarithmic factors, this matches the classical slow prediction rate for ℓ_1-regularized regression without restricted-eigenvalue assumptions; Dalalyan et al. (2017) refine this rate and apply it to one-dimensional TV denoising. Appendix C.6 gives the detailed proof, and Appendix B.4 shows that known upper bounds may replace the exact design and noise scales. The theorem therefore establishes statistical accuracy after finitely many GD iterations.

The effect of k. The order index k enters (5) through $\|D_k \boldsymbol{b}^\star\|_1$, m_k, r_k, and ℓ_k. For the isotropic design $\Sigma = I_p/p$, the prediction scale satisfies $\frac{\mathrm{tr}\{(D_k^\top D_k)^\dagger\}}{pm_k} \le \ell_k^2 \le \frac{1}{p\,\lambda_{\min}^+(D_k^\top D_k)}$, where $\lambda_{\min}^+$ is the smallest positive eigenvalue; Appendix B.3 gives the proof. Choosing k therefore creates a trade-off: matching it to the within-region smoothness can reduce $\|D_k \boldsymbol{b}^\star\|_1$, while increasing it also enlarges the polynomial background and changes the reconstruction geometry. Theorem 3 permits validation over a fixed set of orders with only logarithmic dependence on the number of candidates.

4.3 Validation and the connection to variation penalties

Theorem 2 guarantees accurate iterates within a specified iteration window. The next result shows that an independent validation set can select among saved checkpoints while nearly preserving the risk of the best candidate.

Theorem 3 (Validation selection)**.** *Conditional on training, let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_s$ be s candidates generated without validation data, and let $\widehat{\boldsymbol{b}}$ minimize their validation MSE. Validation data are independent of training and obey Assumptions 1–2 with the same $\boldsymbol{b}^\star, \Sigma, \sigma, \omega$. If $0 < \epsilon < 1/4$ and $n_v \ge C\omega^4\epsilon^{-2}\log(4s/\delta)$, then, with conditional probability at least $1 - \delta$,*

$$\mathcal{R}(\widehat{\boldsymbol{b}}) \le \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 \min_{j \le s} \mathcal{R}(\boldsymbol{b}_j) + \frac{4\sigma^2 \log(4s/\delta)}{\epsilon(1-\epsilon)^2 n_v}. \quad (6)$$

The number of candidates s enters the guarantee in (6) only through $\log s$. The candidate set may include the background-only fit and checkpoints from several different orders k, provided that the

orders and checkpoint locations are fixed before examining validation data. For the run in Theorem 2, saving iterations $1, 2, 4, \ldots$ through T_+ guarantees that one saved iterate lies in the window $[T_-, T_+]$. If $n_v \geq C\omega^4\epsilon^{-2}\log(8s/\delta)$, Theorem 3 transfers the rate in (5) to the selected estimator, up to the displayed multiplicative factor and additive validation term, with joint probability at least $1-\delta$. Appendices C.7, C.8, and B.5 give the checkpoint, validation, and multi-order arguments.

We finally study the continuous-time version of factor GD. The result first lets gradient flow converge for a fixed initialization size α, and then sends α to zero.

Proposition 4 (Gradient-flow variation preference)**.** *Suppose $\gamma > 0$, the interpolation constraint $X\boldsymbol{b} = \boldsymbol{y}$ is feasible, and $\operatorname{rank}(XQ_k) = r_k$. For every $\alpha > 0$, the factor gradient flow initialized by $\boldsymbol{u}(0) = \boldsymbol{v}(0) = \alpha\mathbf{1}_{m_k}$ has a coefficient limit $\boldsymbol{b}_\alpha$. As $\alpha \downarrow 0$, this family is bounded and every accumulation point belongs to $\arg\min_{\boldsymbol{b}:\, X\boldsymbol{b}=\boldsymbol{y}} \|D_k\boldsymbol{b}\|_1$. Uniqueness of this minimum-variation interpolant implies convergence of $\boldsymbol{b}_\alpha$.*

For fixed α, gradient flow selects an interpolating solution through an initialization-dependent objective derived in Appendix C.9. As α approaches zero, this objective favors smaller $\|D_k\boldsymbol{b}\|_1$, which yields the proposition. This parallels initialization bias in diagonal linear networks (Woodworth et al., 2020) and connects the method to TV and higher-order grid variation (Sadhanala et al., 2017). The result concerns converged continuous-time flow followed by vanishing initialization; it does not claim that finite-step GD equals an explicitly penalized estimator.

5 Experiments

We evaluate GD across several simulation settings and a real-data application, comparing its predictive accuracy, coefficient recovery, and computation with competing methods. The simulations include curved and oblique boundaries as well as regions with different degrees of smoothness. We also examine the effects of the consistency weight γ, initialization α, and order index k, and compare finite-step behavior with the theoretical findings in Section 4. Training, validation, and test data are separate; Appendix A gives protocols and supplementary results.

Comparison with competing methods. We generate responses from $y_i = \boldsymbol{x}_i^\top\boldsymbol{b}^\star + \varepsilon_i$, with independent $\boldsymbol{x}_i \sim N(\mathbf{0}, I_p/p)$ and $\varepsilon_i \sim N(0, 0.25^2)$. The four $64 \times 64 \times 64$ coefficient images are a sphere, a cuboid, an oblique split, and a six-region image; each is centered and scaled so that $\|\boldsymbol{b}^\star\|_2^2/p = 1$. These images range from simple regions to non-axis-aligned and multi-region boundaries. This setting is also used for the sensitivity analysis below and, at a smaller grid and sample size, the benchmark in Appendix A.1; that appendix gives the exact image definitions.

At $k = 0$, we compare GD with four baselines. Anisotropic TV solves $\min_{\boldsymbol{b}}\{\|X\boldsymbol{b} - \boldsymbol{y}\|_2^2/(2n) + \lambda\|D_0\boldsymbol{b}\|_1\}$ (Wang & Zhu, 2017). CP regression constrains $\mathcal{B}$ to a low-rank CANDECOMP/PARAFAC decomposition (Zhou et al., 2013). Ridge and elastic net (EN) penalize the vectorized coefficient image by an ℓ_2 penalty and a tuned ℓ_1–ℓ_2 mixture, respectively. Each image has 1024 training, 256 validation, and $n_{\text{test}} = 1024$ test observations. We report coefficient MSE $\|\widehat{\boldsymbol{b}} - \boldsymbol{b}^\star\|_2^2/p$ and test MSE $\|\boldsymbol{y}_{\text{test}} - X_{\text{test}}\widehat{\boldsymbol{b}}\|_2^2/n_{\text{test}}$, together with complete fit-and-selection time. There are 100 paired seeds per image; because CP is substantially more expensive, it uses the first 20. Validation selects GD checkpoints and baseline tuning parameters. GD uses the balanced initialization $\alpha = 10^{-12}$. Within each image–seed case, all methods run serially on an NVIDIA RTX A6000 GPU. The two GD budgets, $T_{\max} = 6000$ and 12000, are timed separately, and the complete TV tuning path is run once per case. Every validation-selected TV fit met the 10^{-5} relative primal–dual gap target; a single unselected candidate did not and was retained in the timing.

Table 1 shows that GD(12000) has the lowest test and coefficient MSE for every image. GD(6000) closely matches or improves on TV while using less than half of its fit-and-selection time; GD(12000) also remains faster. CP is slower and recovers simple shapes more readily than oblique or multi-region structure. Paired comparisons on its first 20 seeds preserve the same overall ranking. Figure 3 in Appendix A.1 shows that Ridge and EN fail to recover the spatial signal and CP misses complex boundaries, whereas GD and TV recover the piecewise-constant regions and their boundaries. Appendix A.1 reports the corresponding smaller-sample benchmark and the full timing protocol; the predictive and computational advantages of GD remain stable.

Table 1: $64 \times 64 \times 64$ benchmark by coefficient image. Entries are mean (SD) over 100 paired seeds per image; CP uses the first 20. Seconds cover complete fitting and validation selection.

Image	GD(6000)	GD(12000)	TV	CP	Ridge	EN
Test MSE						
Sphere	0.1642 (0.0317)	**0.1546** (0.0099)	0.1728 (0.0097)	0.6171 (0.2886)	1.0606 (0.0506)	1.0648 (0.0522)
Cuboid	0.1172 (0.0192)	**0.1106** (0.0087)	0.1572 (0.0115)	0.3351 (0.3362)	1.0629 (0.0518)	1.0664 (0.0513)
Oblique split	0.1155 (0.0057)	**0.1136** (0.0051)	0.1204 (0.0058)	1.0073 (0.1321)	1.0615 (0.0453)	1.0644 (0.0458)
Multi-region	0.1568 (0.0120)	**0.1517** (0.0098)	0.1566 (0.0101)	1.0459 (0.0515)	1.0583 (0.0424)	1.0620 (0.0448)
Coefficient MSE						
Sphere	0.1016 (0.0292)	**0.0922** (0.0073)	0.1107 (0.0068)	0.5570 (0.2834)	0.9972 (0.0014)	1.0011 (0.0083)
Cuboid	0.0556 (0.0211)	**0.0487** (0.0079)	0.0946 (0.0104)	0.2674 (0.3189)	0.9971 (0.0013)	1.0009 (0.0090)
Oblique split	0.0531 (0.0039)	**0.0513** (0.0030)	0.0581 (0.0036)	0.9480 (0.1274)	0.9977 (0.0016)	1.0003 (0.0057)
Multi-region	0.0949 (0.0081)	**0.0897** (0.0058)	0.0947 (0.0057)	0.9974 (0.0117)	0.9974 (0.0015)	1.0014 (0.0093)
Seconds						
Sphere	53.56 (0.11)	104.38 (0.15)	136.07 (4.38)	1149.73 (77.02)	1.45 (0.02)	249.77 (8.88)
Cuboid	53.55 (0.09)	104.38 (0.12)	126.76 (5.26)	1122.67 (72.87)	1.45 (0.02)	249.13 (7.52)
Oblique split	53.49 (0.09)	104.34 (0.13)	139.66 (15.86)	1158.21 (74.68)	1.45 (0.02)	249.44 (8.83)
Multi-region	53.53 (0.11)	104.40 (0.14)	150.86 (4.32)	1175.32 (78.05)	1.45 (0.02)	249.18 (8.22)

Table 2: Coefficient MSE sensitivity to consistency weight γ and initialization α: mean (SD) over 100 paired seeds per coefficient image on a $16 \times 16 \times 16$ grid with $n = 256$, $k = 0$, and 6000 steps.

Coefficient image				
Consistency weight γ	0	$0.25\|\widetilde{X}_0\|_{\mathrm{op}}^2$	$\|\widetilde{X}_0\|_{\mathrm{op}}^2$	$4\|\widetilde{X}_0\|_{\mathrm{op}}^2$
Sphere	0.3645 (0.0368)	0.0995 (0.0237)	0.0940 (0.0230)	0.0958 (0.0225)
Cuboid	0.2920 (0.0339)	0.0207 (0.0108)	0.0164 (0.0096)	0.0153 (0.0092)
Oblique split	0.2117 (0.0184)	0.0781 (0.0094)	0.0761 (0.0094)	0.0758 (0.0090)
Multi-region	0.2356 (0.0180)	0.1244 (0.0136)	0.1243 (0.0136)	0.1265 (0.0139)
Initialization α	10^{-16}	10^{-12}	10^{-8}	10^{-4}
Sphere	0.0918 (0.0229)	0.0940 (0.0230)	0.0988 (0.0232)	0.1167 (0.0246)
Cuboid	0.0151 (0.0092)	0.0164 (0.0096)	0.0194 (0.0105)	0.0334 (0.0148)
Oblique split	0.0756 (0.0092)	0.0761 (0.0094)	0.0771 (0.0094)	0.0815 (0.0094)
Multi-region	0.1239 (0.0135)	0.1243 (0.0136)	0.1251 (0.0134)	0.1284 (0.0128)

Sensitivity analysis for consistency weight γ and initialization α. We retain this setting and the four coefficient images above on a $16 \times 16 \times 16$ grid, with 256 training, 256 validation, and 1024 test observations over 100 paired seeds per image. At $k = 0$, we compare $\gamma \in \{0, 0.25, 1, 4\}\|\widetilde{X}_0\|_{\mathrm{op}}^2$ for GD(6000) and GD(12000), using validation-selected checkpoints. The upper panel of Table 2 reports GD(6000): removing consistency sharply increases coefficient MSE for every image, whereas positive weights perform similarly, supporting a fixed positive weight across these settings.

With $\gamma = \|\widetilde{X}_0\|_{\mathrm{op}}^2$ fixed, we then compare $\alpha \in \{10^{-16}, 10^{-12}, 10^{-8}, 10^{-4}\}$ under the same setting. The lower panel of Table 2 shows similar errors for the two smallest initializations, including the default 10^{-12}, and increasing errors for larger α. Both sensitivity patterns persist under GD(12000); Appendix A.2 gives these results.

Effect of the order index k. Using the same Gaussian design, noise model, and validation–test split, we compare $k = 0, 1, 2$ on five $16 \times 16 \times 16$ coefficient images (Appendix A.3) with 256 training observations. We use 100 paired seeds, the default γ, and a 24000-step budget with validation-based early stopping. Table 3 reports coefficient MSE, while the left panel of Figure 2 shows the corresponding mean coefficient images. The order $k = 0$ favors jumps and curved boundaries, $k = 1$ favors continuous linear pieces, and $k = 2$ favors continuous quadratic pieces. Thus the best order depends on both regional smoothness and boundary jumps. The polynomial background also matters: for quadratic pieces, the $k = 2$ background-only fit has MSE 0.01113, compared with 0.01158 for GD, whereas for linear pieces, $k = 1$ reduces the background-only MSE from 0.05027 to 0.01447.

Finite-step behavior. For these same runs, the right panel of Figure 2 shows the median coefficient MSE paths, their 10th–90th percentile bands, and the median validation-selected stopping step. The preferred order changes with the coefficient image, and several paths worsen after reaching an early minimum, supporting validation-based early stopping. Across 300 matched-order trajectories (sphere/$k = 0$, linear/$k = 1$, and quadratic/$k = 2$), the discrete comparison and Armijo checks had no violations above 10^{-10}, and the largest KL-identity error was 2.54×10^{-13}, consistent with the discrete descent relations used in the finite-step analysis.

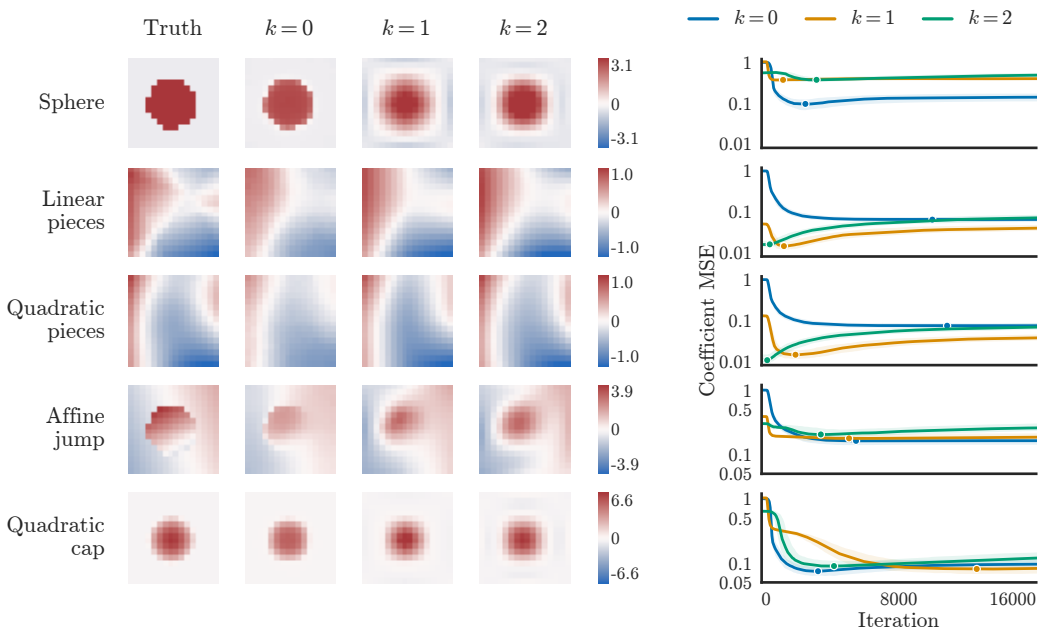


Figure 2: Order index and finite-step behavior. Each row corresponds to one coefficient image. Left panel: truth and voxelwise means of 100 validation-selected GD fits for $k = 0, 1, 2$. Right panel: median coefficient MSE paths, 10th–90th percentile bands, and median validation-selected stopping steps (dots). Paths show steps 0–16000 of 24000 on logarithmic vertical axes.

Table 3: Order-index comparison: mean (SD) coefficient MSE over 100 paired seeds per cell.

Coefficient image	$k = 0$	$k = 1$	$k = 2$
Sphere	**0.0940** (0.0230)	0.3702 (0.0148)	0.3755 (0.0412)
Linear pieces	0.0647 (0.0069)	**0.0145** (0.0026)	0.0161 (0.0027)
Quadratic pieces	0.0751 (0.0079)	0.0146 (0.0028)	**0.0116** (0.0028)
Affine jump	**0.1625** (0.0180)	0.1733 (0.0176)	0.2060 (0.0241)
Quadratic cap	**0.0747** (0.0108)	0.0813 (0.0125)	0.0912 (0.0121)

Table 4: SST results: mean (SD) over 100 training subsamples.

Method	Test MSE	Seconds
GD	**0.1739** (0.0149)	3.09 (0.05)
TV	0.1792 (0.0198)	7.36 (1.26)
CP	0.2260 (0.0533)	8.60 (0.80)
Ridge	0.2266 (0.0347)	0.0112 (0.0003)
EN	0.2251 (0.0545)	18.92 (1.15)

Real-data example. We study the sea surface temperature (SST) data released with Liao et al. (2026). For year i, the predictor $\mathcal{X}_i \in \mathbb{R}^{60\times 50}$ records March–May SST on a two-degree grid with the supplied sea-area weights, and y_i is the log number of category-1-or-higher North Atlantic hurricanes. For each of 100 training subsamples, 40 years sampled from 1950–1994 train the models, 1995–2004 is used for validation, and 2005–2023 is the fixed test block. GD uses $k = 0$ and 6000 steps. All five methods use the same splits and run serially on one AMD EPYC 7543 thread; Appendix A.4 details normalization and timing. Table 4 shows that GD attains the lowest mean test MSE, slightly improving on TV while using less than half its mean runtime. The runs share the same 19 test years and therefore measure training-subsample sensitivity rather than independent test performance. Temporal dependence and future-year shift also limit interpretation.

6 Discussion

We developed a simple GD procedure for piecewise-smooth image regression with finite-step guarantees and competitive predictive and computational performance. Future work could extend the present linear framework to generalized linear and nonlinear models, with corresponding statistical guarantees. Extensions to deep learning could examine how local image structure affects prediction, computation, and spatial interpretability. Incorporating the difference reparameterization into learned image representations could also help preserve spatial structure in nonlinear prediction. Establishing statistical rates for practical backtracking steps is another theoretical direction.

AI USE STATEMENT

In this work, we used generative AI tools for discussing the model framework, hypotheses, methodology, and experimental design, all proposed by the authors; revising the mathematical claims and proof steps, which the authors first formulated; proofreading the proofs, which the authors wrote; implementing the methods and generating synthetic data; and cleaning the real data. We did not use them for qualitative analysis or the interpretation of results. We have reviewed all AI-assisted work: the authors checked every proof, and the AI-assisted code was verified by the authors with operator-identity tests, gradient checks, and an independent convex-solver comparison. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses synthetic images and publicly released sea surface temperature and hurricane-count records. No identifiable human-subject data are used. The real-data analysis is predictive and does not establish a causal climate effect or support operational forecasting.

REPRODUCIBILITY STATEMENT

The model, training procedure, and assumptions are specified in Sections 2–4, and complete proofs appear in Appendix C. All experimental details are given in Appendix A. The anonymized supplementary material contains the code and data needed to reproduce every table and figure directly.

REFERENCES

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. doi: 10.2140/pjm.1966.16.1.

Yang Bai, Ting Li, and Yang Sui. Generalized tensor regression with internal variation regularization. *Statistica Sinica*, 37(2), 2027. doi: 10.5705/ss.202024.0281. Forthcoming.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. doi: 10.1007/s10851-010-0251-1.

Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017. doi: 10.3150/15-BEJ756.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1279. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5da6ce80e97671b70c01a2e703b868b3-Abstract-Conference.html.

Long Feng, Xuan Bi, and Heping Zhang. Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation. *Journal of the American Statistical Association*, 116(533):144–158, 2021. doi: 10.1080/01621459.2020.1766468.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 386–407, 2020. URL https://proceedings.mlr.press/v117/ghai20a.html.

Jeff Goldsmith, Lei Huang, and Ciprian M. Crainiceanu. Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64, 2014. doi: 10.1080/10618600.2012.743437.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1115–1146, 2016. URL https://proceedings.mlr.press/v49/huetter16.html.

Jian Kang, Brian J. Reich, and Ana-Maria Staicu. Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184, 2018. doi: 10.1093/biomet/asx075.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. doi: 10.1137/070690274.

Jiangyuan Li, Thanh V. Nguyen, Chinmay Hegde, and Raymond K. W. Wong. Implicit regularization for group sparsity. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d7Q0vVfJ0wO.

Sijia Liao, Xiaoxiao Sun, Ning Hao, and Hao Helen Zhang. Interpretable scalar-on-image linear regression models via the generalized Dantzig selector. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2690121. Data and code: https://github.com/Sja-88/GDS_Repo.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, volume 2169 of *Lecture Notes in Mathematics*, pp. 277–299. Springer, Cham, 2017. doi: 10.1007/978-3-319-45282-1_18.

Zikai Lin, Junsouk Choi, Ruoxuan Mao, Bangyao Zhao, and Jian Kang. Spatial adaptive selection using binary conditional autoregressive model with application to brain-computer interface. *Journal of Computational and Graphical Statistics*, 35(1):1–12, 2026. doi: 10.1080/10618600.2025.2495256.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16270–16295, 2022. URL https://proceedings.mlr.press/v162/nacson22a.html.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. doi: 10.1145/882262.882269.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/17a9ab4190289f0e1504bbb98d1d111a-Abstract-Conference.html.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. doi: 10.1016/0167-2789(92)90242-F.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James L. Sharpnack, and Ryan J. Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3e60e09c222f206c725385f53d7e567c-Abstract.html.

Yang Sui, Xin He, and Yang Bai. Implicit regularization in over-parameterized support vector machine. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31943–31966, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/659e07806dc17bd69d0d9aed47f85e7c-Abstract-Conference.html.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014. doi: 10.1214/13-AOS1189.

Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011. doi: 10.1214/11-AOS878.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5cf21ce30208cfffaa832c6e44bb567d-Abstract.html.

Xiao Wang and Hongtu Zhu. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017. doi: 10.1080/01621459.2016.1194846.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Sanyou Wu and Long Feng. Sparse Kronecker product decomposition: A general framework of signal region detection in image regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):783–809, 2023. doi: 10.1093/jrsssb/qkad024.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 109(4):1033–1046, 2022. doi: 10.1093/biomet/asac010.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013. doi: 10.1080/01621459.2013.776499.

Appendix A provides supplementary experiments and implementation details. Appendix B develops additional theoretical results, and Appendix C contains proofs of the main results.

A SUPPLEMENTARY EXPERIMENTS

This appendix records simulation and SST protocols, additional numerical results, and implementation details.

A.1 FIRST-DIFFERENCE BENCHMARK

For each grid size $s \in \{32, 64\}$, voxel centers are $(j + 1/2)/s$. Let (x, y, z) be a voxel center. The four unnormalized coefficient images are: the indicator of $(x - .48)^2 + (y - .48)^2 + (z - .48)^2 < .28^2$ (sphere); the indicator that coordinate j lies between $.19 + .025j$ and $.69 + .02j$ for $j = 0, 1, 2$ (cuboid); the indicator of $x + 2y + 3z < 6(.48)$ (oblique split); and a six-region nearest-center image (multi-region). The latter uses centers $(.18, .22, .25), (.75, .2, .3), (.5, .75, .15), (.2, .7, .8), (.8, .75, .75), (.45, .4, .65)$ and assigns respective values $-1, .7, 1.3, -.4, .2, -1.2$. Every true coefficient image is centered and divided by its voxelwise root mean square. The fixed seeds are 20271001–20271100. Within a seed, methods share training, validation, and test data; CP uses the first 20 seeds for each image.

The design has independent $N(0, 1/p)$ entries and response noise has standard deviation 0.25. At $64 \times 64 \times 64$ we use $n = 1024$ and matrix-free predictor products; at $32 \times 32 \times 32$ we use $n = 256$ and a cached design. Both settings use 256 validation and 1024 test observations. GD uses $\alpha = 10^{-12}$ and fixed 6000- and 12000-step budgets, each with validation-selected checkpoints. The other methods follow prespecified training-defined tuning grids. TV uses an anisotropic first-difference penalty and a primal–dual gap stopping target of 10^{-5}. Its selected solution converged in every case; one unselected candidate among the 400 cases at $64 \times 64 \times 64$ did not converge, and its result and time were retained. CP searches ranks $1, 2, 4, 8$ and five penalties with two fixed initializations. All 400 cases per grid passed the saved data and numerical audit.

Within each seed, all methods were run serially on the assigned NVIDIA RTX A6000 GPU; seeds were distributed across four identical GPUs. Timings include method-specific preprocessing, fitting, and validation selection, and exclude common data generation and transfer, warmup, test evaluation, and output serialization. The two GD budgets are separately timed runs; TV has one path per case. Table 5 and Figure 4 give the smaller-grid results. Across the 20 CP-matched seeds per image, GD and TV each have lower mean test MSE than CP for all four images at both sizes, so the ranking is not caused solely by the different seed counts.

Table 5: First-difference benchmark at $32 \times 32 \times 32$ ($n = 256$) by coefficient image. Entries are mean (SD) over 100 paired seeds per image; CP uses the first 20. Seconds cover complete fitting and validation selection.

Image	GD(6000)	GD(12000)	TV	CP	Ridge	EN
Test MSE						
Sphere	0.3473 (0.0859)	**0.3258** (0.0489)	0.3620 (0.0550)	0.9885 (0.1184)	1.0597 (0.0452)	1.0630 (0.0452)
Cuboid	0.3751 (0.1423)	**0.3373** (0.0744)	0.3793 (0.0720)	0.4390 (0.1971)	1.0568 (0.0420)	1.0609 (0.0415)
Oblique split	0.1781 (0.0142)	**0.1779** (0.0140)	0.1855 (0.0155)	1.0737 (0.0610)	1.0487 (0.0467)	1.0531 (0.0464)
Multi-region	0.2541 (0.0213)	**0.2541** (0.0212)	0.2685 (0.0247)	1.0616 (0.0411)	1.0546 (0.0433)	1.0589 (0.0438)
Coefficient MSE						
Sphere	0.2850 (0.0850)	**0.2636** (0.0469)	0.2993 (0.0521)	0.9155 (0.1119)	0.9943 (0.0027)	0.9972 (0.0035)
Cuboid	0.3142 (0.1429)	**0.2766** (0.0751)	0.3182 (0.0729)	0.3774 (0.1911)	0.9947 (0.0028)	0.9985 (0.0108)
Oblique split	0.1163 (0.0119)	**0.1160** (0.0115)	0.1232 (0.0133)	1.0012 (0.0055)	0.9937 (0.0024)	0.9986 (0.0092)
Multi-region	0.1926 (0.0197)	**0.1923** (0.0196)	0.2069 (0.0229)	1.0000 (0.0000)	0.9947 (0.0028)	0.9989 (0.0084)
Seconds						
Sphere	11.67 (0.08)	23.18 (0.15)	23.32 (1.74)	64.55 (8.16)	0.061 (0.006)	8.33 (0.56)
Cuboid	11.68 (0.08)	23.21 (0.15)	24.02 (1.91)	62.66 (7.81)	0.061 (0.006)	8.32 (0.50)
Oblique split	11.59 (0.10)	23.09 (0.18)	21.93 (1.22)	69.46 (8.08)	0.061 (0.006)	8.36 (0.60)
Multi-region	11.63 (0.08)	23.13 (0.16)	24.39 (1.48)	64.75 (5.97)	0.061 (0.006)	8.39 (0.59)

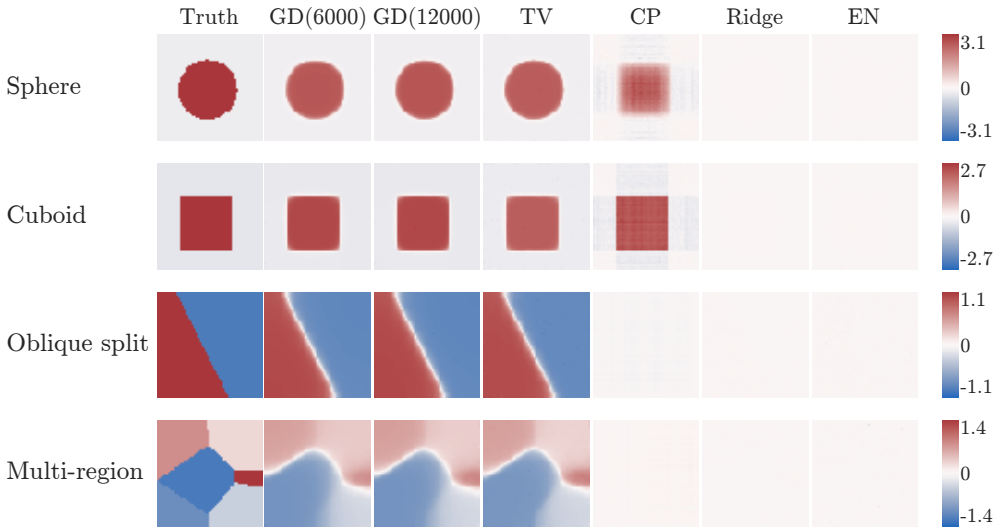


Figure 3: Central slices of the $64 \times 64 \times 64$ coefficient images: truth and voxelwise means of validation-selected fits over 100 seeds, except CP over 20.

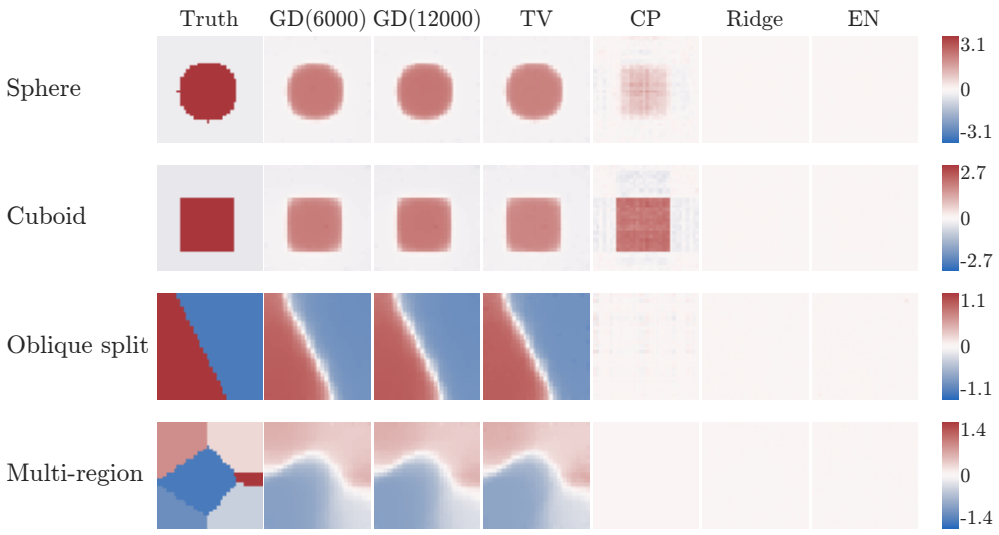


Figure 4: First-difference benchmark: central slices of the $32 \times 32 \times 32$ coefficient images, with the same row and column order as Figure 3. Estimates are voxelwise means of validation-selected fits over 100 seeds, except CP over 20.

A.2 Sensitivity analyses for consistency weight γ and initialization α

We run $k = 0$ GD on a $16 \times 16 \times 16$ grid with 256 training, 256 validation, and 1024 test observations for each of the four benchmark coefficient images. Each image uses 100 paired seeds across parameter values and iteration budgets, and validation-based early stopping selects a checkpoint within each budget. For the consistency-weight analysis, γ is 0, 0.25, 1, or 4 times $\|\widetilde{X}_0\|_{\mathrm{op}}^2$, with $\alpha = 10^{-12}$

fixed. For the initialization analysis, we fix $\gamma = \|\widetilde{X}_0\|_{\mathrm{op}}^2$ and vary $\alpha \in \{10^{-16}, 10^{-12}, 10^{-8}, 10^{-4}\}$ using the same data and checkpoint schedules. Table 6 complements the GD(6000) results in the main text with both sets of GD(12000) coefficient errors. The two smallest initializations remain close, whereas larger values increase error across all four coefficient images.

Table 6: GD(12000) coefficient MSE sensitivity to consistency weight γ and initialization α: mean (SD) over 100 paired seeds per coefficient image.

Consistency weight γ	0	$0.25\|\widetilde{X}_0\|_{\mathrm{op}}^2$	$\|\widetilde{X}_0\|_{\mathrm{op}}^2$	$4\|\widetilde{X}_0\|_{\mathrm{op}}^2$
Sphere	0.3636 (0.0367)	0.0995 (0.0237)	0.0940 (0.0230)	0.0921 (0.0230)
Cuboid	0.2916 (0.0340)	0.0207 (0.0108)	0.0164 (0.0096)	0.0153 (0.0091)
Oblique split	0.2086 (0.0188)	0.0782 (0.0094)	0.0761 (0.0094)	0.0756 (0.0090)
Multi-region	0.2336 (0.0172)	0.1247 (0.0137)	0.1245 (0.0137)	0.1246 (0.0131)
Initialization α	10^{-16}	10^{-12}	10^{-8}	10^{-4}
Sphere	0.0918 (0.0229)	0.0940 (0.0230)	0.0988 (0.0232)	0.1167 (0.0246)
Cuboid	0.0151 (0.0092)	0.0164 (0.0096)	0.0194 (0.0105)	0.0334 (0.0148)
Oblique split	0.0757 (0.0092)	0.0761 (0.0094)	0.0771 (0.0094)	0.0815 (0.0094)
Multi-region	0.1243 (0.0139)	0.1245 (0.0137)	0.1251 (0.0134)	0.1284 (0.0128)

The test MSE and checkpoint-selection frequencies in Table 7 pool all 400 image–seed cases equally for each weight and budget. The boundary fraction records only whether validation selected the last available checkpoint; it does not diagnose convergence. The default weight is nearly unchanged by the larger budget, whereas the fourfold weight selects the 6000-step boundary more often and improves slightly when the budget is extended.

Table 7: Pooled consistency-weight analysis over four coefficient images and 100 seeds per image. Columns give $\gamma/\|\widetilde{X}_0\|_{\mathrm{op}}^2$. Boundary is the fraction of validation-selected checkpoints at the iteration budget.

Budget	Metric	0	0.25	1	4
6000	Test MSE	0.3399	0.1435	0.1405	0.1412
	Boundary (%)	35.00	2.25	0.75	44.00
12000	Test MSE	0.3383	0.1435	0.1405	0.1397
	Boundary (%)	28.25	0.25	0.50	3.25

A.3 Order indices and path diagnostics

The order-index comparison uses a $16 \times 16 \times 16$ grid, $n = 256$, 256 validation observations, 1024 test observations, and the 100 paired seeds 20271001–20271100 for each of five coefficient images and order indices $k = 0, 1, 2$. The design and fixed noise standard deviation match the first-difference benchmark. GD starts from $\alpha = 10^{-12}$, fixes $\gamma = \|\widetilde{X}_k\|_{\mathrm{op}}^2$, runs up to 24000 iterations, and uses validation-based early stopping. The background-only fit is included in the candidate pool. Checkpoints occur every 25 steps through 4000 and at predetermined logarithmically spaced later steps. No step budget was extended after examining these formal results.

The sphere, affine jump, and quadratic cap use the formulas in the source package; for clarity, the jump equals $.25+2(x-.48)-(y-.48)+1.5(z-.48)$ inside the radius-.28 sphere and $-.15+.25(x-.48)+.7(y-.48)-.2(z-.48)$ outside. The cap is $[.28^2-(x-.48)^2-(y-.48)^2-(z-.48)^2]_+$. For the linear and quadratic pieces, put $\boldsymbol{r} = (x, y, z) - (.5, .5, .5)$ and

$$f_a(\boldsymbol{r}) = \sum_{j=1}^{5} c_j [\boldsymbol{d}_j^\top \boldsymbol{r} - t_j]_+^a, \qquad a = 1, 2,$$

where $\boldsymbol{d}_j$ are the normalized rows of $(1,2,3), (-2,1,2), (2,-3,1), (3,1,-2), (-1,-2,3)$, $\boldsymbol{c} = (1, -.8, .6, -1.1, .9)$, and $\boldsymbol{t} = (-.15, .08, -.05, .18, -.12)$. All five raw images are centered and scaled to voxelwise mean square one. Between-seed SDs of the 15 coefficient-MSE cells range from 0.00260 to 0.04119; the full cellwise results and background errors are preserved in the supplementary CSV.

The iteration-path analysis reuses the 100 paths for each of the three order indices and five coefficient images in the order-index comparison (1500 trajectories). The figure summarizes all 15 cells through step 16000 of the 24000-step budget. Finite-step identity diagnostics were saved only for the 300

trajectories on sphere/$k = 0$, linear pieces/$k = 1$, and quadratic pieces/$k = 2$. At their checkpoints, none exceeded 10^{-10} in the exact finite-step comparison or Armijo excess after its stated tolerance. The smallest factor was 2.31×10^{-29} and the minimum positivity margin was 0.553; the maximum absolute logarithmic identity discrepancy was 2.54×10^{-13}.

A.4 SST PROTOCOL

We use the public spring sea surface temperature (SST) images and log North Atlantic hurricane counts released with Liao et al. (2026). The images retain the 60×50 two-degree grid and sea-area weights, and the intercept is unpenalized. Each of 100 prespecified seeds samples 40 training years from 1950–1994. The years 1995–2004 are reserved for validation-based tuning, while the fixed 2005–2023 block is used only for test prediction. Thus the 100 repetitions vary the training subset but share the same validation and test years.

GD uses $k = 0$ and a 6000-step budget. GD, TV, CP, Ridge, and EN use the same split and squared loss for each seed. Normalization is estimated from that seed's training years and then applied to validation and test data. The five fits run sequentially on one AMD EPYC 7543 thread. Reported times include method-specific preprocessing, candidate fitting, validation selection, and test prediction; they exclude file I/O. The test MSE and complete workflow time in Table 4 are summarized across the 100 training subsamples, not across independent test blocks.

A.5 IMPLEMENTATION AND COMPUTATIONAL DETAILS

Let $P_k^{\perp} = I_{m_k} - D_k D_k^{\dagger}$, which projects difference coordinates onto their inconsistent component. The initial and maximum step size is $0.05/\max\{\|\widetilde{X}_k^{\top}\widetilde{\boldsymbol{y}}_k\|_{\infty}, 10^{-12}\}$. Rejected steps are halved; accepted steps increase by 1.02 up to the cap. The implementation permits loss tolerance $10^{-14}\max\{1, \mathcal{L}_k(\boldsymbol{w}_t)\}$ and at most 60 backtracking trials. Each trial requires finite loss and positive factor multipliers. If no trial is accepted within the cap, the run terminates with a reported failure; an unsuccessful trial is never taken as an update. The general path bounds concern successfully executed prefixes. The conservative step and budget in Appendix C.6 are distinct from this experimental configuration. Exact comparison uses monotonicity; (34) supplies the acceptance-tolerance correction. The checkpoint schedule is fixed before inspecting validation values. Storing only the running winner is equivalent to minimizing over the full candidate pool; s in Theorem 3 counts that full pool. The order-zero branch retains the original padded difference layout for exact reproduction. Unused boundary coordinates remain zero; deleting them recovers the valid theoretical operator. Any KL initialization constant in padded coordinates must use the actual dimension or first remove redundant coordinates.

An equivalent loss-difference calculation. For a proposed step, direct expansion gives

$$\begin{aligned} \Delta\boldsymbol{w}_t &= -4\eta_t(\boldsymbol{u}_t^{\odot 2} + \boldsymbol{v}_t^{\odot 2}) \odot \boldsymbol{g}_t + 4\eta_t^2 \boldsymbol{w}_t \odot \boldsymbol{g}_t^{\odot 2}, \\ \mathcal{L}_k(\boldsymbol{w}_t + \Delta\boldsymbol{w}_t) - \mathcal{L}_k(\boldsymbol{w}_t) &= \boldsymbol{g}_t^{\top}\Delta\boldsymbol{w}_t + \tfrac{1}{2}\|\widetilde{X}_k\Delta\boldsymbol{w}_t\|_2^2 + \tfrac{\gamma}{2}\|P_k^{\perp}\Delta\boldsymbol{w}_t\|_2^2. \end{aligned} \tag{7}$$

This algebraically equivalent test can avoid subtracting two nearly equal loss values; the stored experiments used the direct loss test with the stated tolerance.

One-dimensional eigendecompositions cost $O(\sum_j p_j^3)$ once. Spectral transforms cost $O(p\sum_j p_j)$ and differences cost $O(d(k+1)p)$ per application. Background storage costs $O(pr_k)$; coefficient restoration can cost $O(pr_k + np + nr_k)$. Cached and matrix-free designs have different memory requirements.

For the explicit $k = 0$ TV benchmark, the ADMM solver profiles the constant background and uses a Woodbury system for its coefficient update. It forms an $n \times n$ Gram matrix once. Beyond the common eigendecompositions, this setup costs $O(np\sum_j p_j + n^2p + n^3)$: the terms come from transforming n design rows, forming the Gram matrix, and factoring it. With a fixed ADMM penalty, each update costs $O(np + n^2 + p\sum_j p_j + dp)$ for design products, triangular solves, a spectral inverse, and local differences. If the penalty is rebalanced, an additional $O(n^3)$ factorization is required. Dual-gap checks also add work at their scheduled iterations. The 31 candidate TV weights share this setup and use warm starts; their cumulative update cost is proportional to the sum of ADMM iterations over those candidates. GD likewise pays for operator setup and its design-based

choice of γ, so the timing tables compare complete preprocessing, fitting, and selection rather than isolated update costs. A fast DCT implementation at $k = 0$ could lower the spectral-transform term for either method, but was not used in the reported timing runs.

B SUPPLEMENTARY THEORETICAL RESULTS

We collect consistency-weight analysis, grid geometry, selected-estimator bounds, and extensions of the finite-step argument before giving the main proofs in Appendix C. In the appendix, write $c_A = 10^{-4}$, so $4c_A = 0.0004$ in (4).

B.1 THE ROLE OF THE CONSISTENCY WEIGHT

Write $P_k^\perp = I_{m_k} - D_k D_k^\dagger$ for the difference-consistency projection. Since $\widetilde{X}_k P_k^\perp = 0$,

$$\mathcal{L}_k(\boldsymbol{w}) = \tfrac{1}{2}\|\widetilde{X}_k(I_{m_k} - P_k^\perp)\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k\|_2^2 + \tfrac{\gamma}{2}\|P_k^\perp \boldsymbol{w}\|_2^2, \qquad \nabla^2 \mathcal{L}_k = \widetilde{X}_k^\top \widetilde{X}_k + \gamma P_k^\perp.$$

If $P_k^\perp \neq 0$, the largest eigenvalue is $\max\{\|\widetilde{X}_k\|_{\mathrm{op}}^2, \gamma\}$. Matching these scales explains the default but does not make the Hessian nonsingular: consistent directions may still be unobserved. In one dimension D_k has full row rank, so $P_k^\perp = 0$ and the consistency weight has no effect.

For the factor objective f, the Hessian is

$$\nabla^2 f = J^\top(\widetilde{X}_k^\top \widetilde{X}_k + \gamma P_k^\perp)J + \begin{bmatrix} 2\,\mathrm{diag}(\boldsymbol{g}) & 0 \\ 0 & -2\,\mathrm{diag}(\boldsymbol{g}) \end{bmatrix}, \qquad J = [2\,\mathrm{diag}(\boldsymbol{u}), -2\,\mathrm{diag}(\boldsymbol{v})]. \tag{8}$$

It also depends on the factors and residual gradient, so the loss Hessian alone does not give a universal factor-GD step rule. With $S_t = \mathrm{diag}(\boldsymbol{u}_t^{\odot 2} + \boldsymbol{v}_t^{\odot 2})$, continuous dynamics satisfy

$$\frac{d(P_k^\perp \boldsymbol{w}_t)}{dt} = -4P_k^\perp S_t \widetilde{X}_k^\top(\widetilde{X}_k \boldsymbol{w}_t - \widetilde{\boldsymbol{y}}_k) - 4\gamma P_k^\perp S_t P_k^\perp \boldsymbol{w}_t.$$

Generally $P_k^\perp S_t \widetilde{X}_k^\top \neq 0$: difference inconsistency need not decay independently or monotonically. Under the interpolation-feasibility conditions of Proposition 4, positive γ does not change the ideal endpoint but can change the early-stopped estimate. When $\gamma = 0$, the interpolating flow instead minimizes the same potential over $\widetilde{X}_k \boldsymbol{w} = \widetilde{\boldsymbol{y}}_k$ without requiring difference consistency $P_k^\perp \boldsymbol{w} = 0$. Its vanishing-initialization limit favors $\|\boldsymbol{w}\|_1$ over unconstrained difference coordinates, which need not equal minimum $\|D_k \boldsymbol{b}\|_1$: the projection $I_{m_k} - P_k^\perp$ is not generally an ℓ_1 contraction.

Dividing the objective by $\|\widetilde{X}_k\|_{\mathrm{op}}^2 > 0$ gives unit consistency weight under the default. Identical discrete iterates also require multiplying step sizes by $\|\widetilde{X}_k\|_{\mathrm{op}}^2$ and rescaling step caps and absolute tolerances. The implementation retains the original scale. If $\widetilde{X}_k = 0$, balanced initialization gives $\boldsymbol{w} = \boldsymbol{g} = \boldsymbol{0}$; the algorithm returns the background without division by zero.

B.2 NONPERIODIC GEOMETRY AND A NONTRIVIAL HIGH-DIMENSIONAL REGIME

In this subsection, fix $k = 0$, $q \geq 2$, a $q \times q \times q$ grid, and $\Sigma = I_p/p$ with $p = q^3$. The one-dimensional valid-difference normal matrix is the path Laplacian. Its orthonormal eigenvectors and eigenvalues are

$$\phi_0(j) = q^{-1/2}, \qquad \phi_a(j) = \sqrt{2/q}\cos\{\pi a(j - \tfrac{1}{2})/q\}, \qquad \lambda_a = 4\sin^2\{\pi a/(2q)\}, \quad 1 \leq a < q.$$

Substitution verifies both interior and endpoint equations. Product eigenvectors indexed by $\boldsymbol{a} = (a_1, a_2, a_3)$ have eigenvalues $\lambda_{\boldsymbol{a}} = \lambda_{a_1} + \lambda_{a_2} + \lambda_{a_3}$. For an edge indexed by j in direction s, the cosine difference identity gives $|(D_0\phi_{\boldsymbol{a}})_j|^2 \leq 8\lambda_{a_s}/q^3$. Since $D_0^\dagger = (D_0^\top D_0)^\dagger D_0^\top$,

$$\|D_0^\dagger \boldsymbol{e}_j\|_2^2 = \sum_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{|(D_0\phi_{\boldsymbol{a}})_j|^2}{\lambda_{\boldsymbol{a}}^2} \leq \frac{8}{q^3}\sum_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{1}{\lambda_{\boldsymbol{a}}}.$$

The inequality $\sin x \geq 2x/\pi$ for $0 \leq x \leq \pi/2$ implies $\lambda_{\boldsymbol{a}} \geq 4\|\boldsymbol{a}\|_2^2/q^2$. The shell $\max_s a_s = b$ contains at most $(b+1)^3 - b^3 \leq 7b^2$ indices, so

$$\sum_{\boldsymbol{a} \neq \boldsymbol{0}} \lambda_{\boldsymbol{a}}^{-1} \leq \frac{q^2}{4}\sum_{b=1}^{q-1} \frac{7b^2}{b^2} \leq \tfrac{7}{4}q^3, \qquad \ell_0^2 \leq 14/q^3.$$

For the lower bound, there are $p-1$ nonzero eigenvalues, all at most 12. The largest squared column norm is at least the mean, and $m_0 \le 3p$. Hence

$$\ell_0^2 \ge \frac{\mathrm{tr}\{(D_0^\top D_0)^\dagger\}}{pm_0} \ge \frac{p-1}{12pm_0} \ge \frac{1}{72p}.$$

Now take $n = q^2$, $V = C_V q^2$ for fixed $C_V > 0$, and $\alpha = q^{-10}$. The variation class includes fixed-amplitude, fixed-proportion block indicators after centering and normalization. One has $W_\alpha = O(q^2 \log q)$, and the three terms of (5) are respectively

$$O\{\sigma q^{-1/2}(\log q)^{3/2}\}, \qquad O\{q^{-1}(\log q)^3\}, \qquad O(\sigma^2 q^{-2})$$

for fixed confidence and ω. The sample condition holds for sufficiently large q even though $p/n = q \to \infty$. In this example, the risk equals $\|\mathcal{B} - \mathcal{B}^\star\|_F^2/p$, not the unnormalized squared error.

For fixed positive σ, the proof-local scales ρ in (53) and β, Λ, G in (54) satisfy $\beta \asymp 1$, $\Lambda \asymp q$, $G^2 \asymp q$, and $\rho \asymp q^{-5/2}\sqrt{\log q}$. When $\sigma = 0$, instead $\rho \asymp q^{-3}(\log q)^2$. Equation (57) consequently gives the loose polynomial budgets $T_+ = O(q^6)$ and $O(q^7)$ respectively. These estimates show finite attainability, not practical efficiency. The spectral example is specific to first differences; higher orders require their own geometric scaling.

B.3 Order and scale dependence

The prediction scale ℓ_k is related to the pseudoinverse column norms used in TV denoising theory (Hütter & Rigollet, 2016). For $\Sigma = I_p/p$, it admits the following finite-grid bounds, with no periodic boundary assumption:

Lemma B.1 (Spectral bounds for the prediction scale)**.** *With the valid difference operator in* (1) *and* $\Sigma = I_p/p$*,*

$$\frac{\mathrm{tr}\{(D_k^\top D_k)^\dagger\}}{pm_k} \le \ell_k^2 \le \frac{1}{p\,\lambda_{\min}^+(D_k^\top D_k)}, \tag{9}$$

where $\lambda_{\min}^+$ *denotes the smallest strictly positive eigenvalue.*

Proof of Lemma B.1. By definition, $p\ell_k^2 = \max_j \|D_k^\dagger \boldsymbol{e}_j\|_2^2$. The maximum column norm squared is at least its average, so

$$p\ell_k^2 \ge \frac{\|D_k^\dagger\|_F^2}{m_k} = \frac{\mathrm{tr}\{(D_k^\top D_k)^\dagger\}}{m_k}.$$

Each column norm is at most the operator norm. The nonzero singular values of the pseudoinverse are the reciprocals of those of D_k, giving

$$p\ell_k^2 \le \|D_k^\dagger\|_{\mathrm{op}}^2 = \frac{1}{\lambda_{\min}^+(D_k^\top D_k)}.$$

□

The upper bound may be loose. Replacing D_k by cD_k, $c > 0$, multiplies $\|D_k \boldsymbol{h}\|_1$ by c and divides ℓ_k by c. Their product is unchanged for a fixed image error, although training with fixed numerical initialization need not be invariant. The sample condition in Theorem 1 controls the curvature event; small risk still depends on these geometric quantities and the trajectory remainder.

B.4 Unknown scales with known upper bounds

Suppose known bounds satisfy $\|D_k \boldsymbol{b}^\star\|_1 \le V_{\max}$, $\sigma \le \sigma_{\max}$, and $\ell_k \le \bar{\ell}$. They may come from a prescribed model class; they are not inferred from the validation responses. Let $W_{\max}$ be W_α from Section 4.2 with $V = V_{\max}$, and define $\beta_{\max}, \Lambda_{\max}, G_{\max}$ by (54) using $\bar{\ell}, V_{\max}, \sigma_{\max}$. Choose a deterministic $\rho_0 > 0$ and put

$$\begin{aligned}
F_{\max} &= 4\sigma_{\max}^2\{\log 5 + \log(20/\delta)/n\},\\
W_{\mathrm{run}} &= W_{\max} + F_{\max}/\rho_0,\\
\rho_{\max} &= \rho_0 + 2\sqrt{2}\,\sigma_{\max}\bar{\ell}\sqrt{\log(20m_k/\delta)/n} + \omega^4\bar{\ell}^2 W_{\max}\log(2m_k)/n.
\end{aligned}$$

Set the initial step and cap to

$$\eta = \min\left\{\frac{1}{4G_{\max}}, \frac{\rho_0}{2G_{\max}^2}, \frac{1-c_A}{48\beta_{\max}W_{\mathrm{run}} + G_{\max}}, \frac{1}{64\rho_{\max}}\right\}, \tag{10}$$

with the same zero-denominator convention. Run to the predetermined $T_{\max} = \lfloor 1/(16\rho_0\eta)\rfloor$, retaining the background and powers-of-two checkpoints. For this argument, put $\boldsymbol{w}^\star = D_k\boldsymbol{b}^\star$, $\vartheta = \|\boldsymbol{w}^\star\|_1$, $\boldsymbol{\xi} = \Pi_k\boldsymbol{\varepsilon}/\sqrt{n}$, and $M_t = \sum_e(u_{t,e}^2 + v_{t,e}^2)$. Define $w_e^{\star+} = \max\{w_e^\star, 0\}$ and $w_e^{\star-} = \max\{-w_e^\star, 0\}$, and use the generalized KL potential

$$\mathcal{D}_t = \sum_e \left[w_e^{\star+}\log\frac{w_e^{\star+}}{u_{t,e}^2} - w_e^{\star+} + u_{t,e}^2 + w_e^{\star-}\log\frac{w_e^{\star-}}{v_{t,e}^2} - w_e^{\star-} + v_{t,e}^2\right],$$

with $0\log 0 = 0$. This is (28) with comparator $\boldsymbol{w}^\star$.

Execution of the entire prescribed run. On the training noise event (70), $\mathcal{L}_k(\boldsymbol{w}^\star) = \|\boldsymbol{\xi}\|_2^2/2 \le F_{\max}$. Convexity and nonnegativity of the loss imply $\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle \ge -\mathcal{L}_k(\boldsymbol{w}^\star) \ge -F_{\max}$. Set $Z_t = \mathcal{D}_t + 2V_{\max}$; its initial value is at most $W_{\max}$. Using the mass bound (60) and exact identity (30),

$$Z_{t+1} \le Z_t + 4\eta F_{\max} + 4\eta^2 G_{\max}^2(M_t + 2\vartheta) \le (1 + 8\eta^2 G_{\max}^2)Z_t + 4\eta F_{\max}.$$

Therefore, writing $s_t = t\eta$,

$$M_t \le 2Z_t \le 2e^{8\eta G_{\max}^2 s_t}(W_{\max} + 4F_{\max}s_t) \le 2e^{1/4}\left(W_{\max} + \frac{F_{\max}}{4\rho_0}\right) < 3W_{\mathrm{run}}$$

for $s_t \le 1/(16\rho_0)$. Repeating the Hessian bound (62) and the joint induction leading to (65) with the larger envelopes proves first-trial acceptance throughout this longer run.

Coverage of an unknown good window. For analysis only, put $\vartheta = \|D_k\boldsymbol{b}^\star\|_1$, let $W_\alpha(\vartheta)$ denote W_α from Section 4.2 with $V = \vartheta$, and define

$$\rho_{\mathrm{true}} = \rho_0 + 2\sqrt{2}\,\sigma\ell_k\sqrt{\log(20m_k/\delta)/n} + \omega^4\ell_k^2 W_\alpha(\vartheta)\log(2m_k)/n.$$

Then $\rho_0 \le \rho_{\mathrm{true}} \le \rho_{\max}$. Let G_{true} be the gradient envelope in (54) evaluated at $V = \vartheta$ and the true σ, ℓ_k. Equation (10) satisfies all the small-step conditions used for the true-scale mass recursion: in particular $\eta G_{\mathrm{true}}^2 \le \rho_{\mathrm{true}}/2$, $\eta \le 1/(64\rho_{\mathrm{true}})$, and its Hessian cap is no larger than the true-scale cap. Repeating the proof of Theorem 2, from the factor-mass bound (60) through the risk bound (5), up to this shorter window gives the same rate with $W_\alpha(\vartheta)$ and the additional term $CW_\alpha(\vartheta)\rho_0$. The integer window

$$\left\lceil \frac{1}{32\rho_{\mathrm{true}}\eta} \right\rceil \le T \le \left\lfloor \frac{1}{16\rho_{\mathrm{true}}\eta} \right\rfloor$$

is contained in the prescribed run and contains a powers-of-two checkpoint. Apply Theorem 3 with $s = 2 + \lfloor \log_2 T_{\max}\rfloor$ to obtain the corresponding selected-estimator guarantee, including the validation term. The true scales in this paragraph are used only in the proof; the run uses their known upper bounds and ρ_0. This is not a tuning-free guarantee: an arbitrarily small ρ_0 can make the step and budget extremely conservative. If only a covariance-class bound $\Sigma \preceq CI_p/p$ is known, a valid $\bar{\ell}$ is $\sqrt{C/p}\max_j\|D_k^\dagger\boldsymbol{e}_j\|_2$.

B.5 A BOUND FOR THE VALIDATION-SELECTED ESTIMATOR

The following consequence connects training and the returned estimator. Its additional notation is local to this subsection. Fix a finite collection $\mathcal{K}$ of valid order indices before observing validation data. For index k, let $\mathcal{T}_k$ be its predetermined finite set of positive checkpoints and write $\boldsymbol{b}_{k,T}$ for its candidate; $\boldsymbol{b}_{k,0} = \boldsymbol{b}(\boldsymbol{0})$ is its background-only fit. Paths, training-based rank checks, and numerical termination decisions must be independent of validation. Let

$$s = \sum_{k\in\mathcal{K}}(1 + |\mathcal{T}_k|)$$

upper-bound the complete candidate pool, even if some orders or checkpoints are not executed. Minima below are over candidates actually generated.

For $T \geq 1$, put $\boldsymbol{h}_{k,T} = \boldsymbol{b}_{k,T} - \boldsymbol{b}^\star$ locally. For each k and failure level $a \in (0,1)$, use

$$\nu_a = \sigma \max_{j \leq m_k} \|\widetilde{X}_k \boldsymbol{e}_j\|_2 \sqrt{\frac{2\log(4m_k/a)}{n}}, \qquad \kappa = \tfrac{1}{4}, \qquad \tau = \frac{c_1 \omega^4 \ell_k^2 \log(2m_k)}{n}.$$

These are the coordinate-noise and curvature scales in Lemma C.3 and (47); c_1 is the absolute constant of Lemma C.4. With all quantities evaluated at order k, and $\mathcal{E}_T$ as in (26), define

$$B_{k,T} = \frac{1}{\kappa}\Bigg[2\nu_{\delta/(4|\mathcal{K}|)}\|D_k \boldsymbol{h}_{k,T}\|_1 + 2\mathcal{E}_T(D_k \boldsymbol{b}^\star) + \tau\|D_k \boldsymbol{h}_{k,T}\|_1^2 + \frac{8\sigma^2}{n}\left\{r_k \log 5 + \log\frac{16|\mathcal{K}|}{\delta}\right\}\Bigg], \tag{11}$$

$$B_{k,0} = \frac{1}{\kappa}\left[\|\widetilde{X}_k D_k \boldsymbol{b}^\star\|_2^2 + \tau\|D_k \boldsymbol{b}^\star\|_1^2 + \frac{8\sigma^2}{n}\left\{r_k \log 5 + \log\frac{16|\mathcal{K}|}{\delta}\right\}\right]. \tag{12}$$

Under acceptance slack, replace $\mathcal{E}_T$ by the corrected remainder in (34). The separate expression at $T = 0$ avoids division by $s_0 = 0$.

Corollary B.2 (Training and validation together)**.** *Under the training assumptions of Theorem 1 and the validation assumptions of Theorem 3, suppose*

$$n \geq c_0 \omega^4 \max_{k \in \mathcal{K}}\left\{r_k + \log\frac{32|\mathcal{K}|m_k}{\delta}\right\}, \qquad n_v \geq c\omega^4 \epsilon^{-2}\log(8s/\delta).$$

Then, with probability at least $1-\delta$, the validation-selected candidate satisfies

$$\mathcal{R}(\widehat{\boldsymbol{b}}) \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 \min_{k,T} B_{k,T} + \frac{4\sigma^2 \log(8s/\delta)}{\epsilon(1-\epsilon)^2 n_v}. \tag{13}$$

The same event bounds coefficient error by this right side divided by $\lambda_{\min}(\Sigma)$.

Proof of Corollary B.2. For each order, allocate failure probability $\delta/(4|\mathcal{K}|)$ to curvature and $\delta/(4|\mathcal{K}|)$ to Lemma C.3. On their intersection, (51) with the corresponding noise level gives $\mathcal{R}(\boldsymbol{b}_{k,T}) \leq B_{k,T}$ simultaneously for every executed $T \geq 1$. For the background, $D_k \boldsymbol{b}_{k,0} = \boldsymbol{0}$. Curvature and (44), using the same background-noise event, give

$$\kappa\mathcal{R}(\boldsymbol{b}_{k,0}) \leq \|\widetilde{X}_k D_k \boldsymbol{b}^\star\|_2^2 + \frac{\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2}{n} + \tau\|D_k \boldsymbol{b}^\star\|_1^2 \leq \kappa B_{k,0}.$$

A union bound over orders costs at most $\delta/2$, with no union bound over training times. Conditional on training, apply Theorem 3 at level $\delta/2$, using s as an upper bound on pool size. Combining these events and minimizing the training upper bounds proves (13). □

This guarantee covers the estimator returned by a fixed protocol.

B.6 Boundary cases and numerical acceptance in Theorem 2

If $\sigma = 0$, the second term of (53) remains positive, and the statistical window is finite. If $D_k \boldsymbol{b}^\star = \boldsymbol{0}$ and $\sigma = 0$, the initialized gradient vanishes and the identifiable background is exactly the truth. If $\alpha = 0$, all factors remain zero, so a nonbackground truth cannot be learned. For positive initialization the logarithmic dependence in W_α is retained; the bound does not improve monotonically as initialization shrinks.

Theorem 2 uses the positive design-based consistency weight. Without positive γ, the interpolation preference need not minimize $\|D_k \boldsymbol{b}\|_1$; Proposition 4 also requires $\gamma > 0$. Ill-conditioned Σ enters both ℓ_k and the coefficient-error conversion and is not hidden in an absolute constant.

For the conservative cap in (55), exact arithmetic already gives first-trial exact Armijo acceptance. Allowing a nonnegative acceptance slack therefore leaves this mathematical trajectory unchanged. More generally, if only relaxed acceptance is known, let $\mu_t = a\max\{1, \mathcal{L}_k(\boldsymbol{w}_t)\}$ with $a = 10^{-14}$.

It implies $\max\{1, \mathcal{L}_k(\boldsymbol{w}_t)\} \le (1+a)^t \max\{1, \Lambda\}$ whenever the initial loss is bounded by Λ. Thus the comparison correction obeys

$$\frac{1}{s_T} \sum_{t<T} \eta_t \sum_{j=t}^{T-1} \mu_j \le \sum_{j<T} \mu_j \le \max\{1, \Lambda\}\{(1+a)^T - 1\}. \tag{14}$$

For very large theoretical budgets, (14) need not be small, so the experimental configuration is covered by the general executed-prefix bounds with their acceptance correction rather than by Theorem 2.

C Proofs of the main results

Fix an order index k throughout each proof. All vectors use the same notation as in the main text; scalar entries remain unbolded. Unless stated otherwise, constants denoted by c, C, and C_0 are positive absolute constants that may change between displays; c_0, c_1, and C_{md} denote the fixed constants of Lemma C.4. For this appendix, write $\mathrm{Pred}_n(\boldsymbol{b}) = \|X(\boldsymbol{b} - \boldsymbol{b}^\star)\|_2^2/n$ for empirical prediction error. To keep repeated projection products readable, write

$$P_k^\perp = I_{m_k} - D_k D_k^\dagger. \tag{15}$$

This projects difference coordinates onto their inconsistent component, whereas Π_k projects in sample space.

C.1 Difference geometry and the background

Let $D_j^{(k+1)}$ be the $(p_j - k - 1) \times p_j$ matrix of valid one-dimensional differences. Its null space consists of polynomial sequences of degree at most k: the difference recurrence determines every later value from the first $k+1$ entries, and the $k+1$ independent monomials are annihilated. In particular, the matrix has full row rank. For the stacked grid operator,

$$D_k^\top D_k = \sum_{j=1}^{d} \left(\bigotimes_{\ell<j} I_{p_\ell} \right) \otimes \{(D_j^{(k+1)})^\top D_j^{(k+1)}\} \otimes \left(\bigotimes_{\ell>j} I_{p_\ell} \right). \tag{16}$$

An empty Kronecker product is the scalar 1. In the product eigenbasis, all eigenvalues are sums of nonnegative one-dimensional eigenvalues. A sum is zero precisely when every summand is zero. Thus

$$\ker D_k = \operatorname{span}\left\{ \prod_{j=1}^{d} x_j^{a_j} : 0 \le a_j \le k \right\}, \qquad r_k = \dim \ker D_k = (k+1)^d. \tag{17}$$

The degree restriction is coordinatewise, not a restriction to total degree k.

The Moore–Penrose identities give

$$\begin{aligned} D_k D_k^\dagger D_k &= D_k, & D_k^\dagger D_k D_k^\dagger &= D_k^\dagger, \\ D_k^\dagger D_k &= I_p - Q_k Q_k^\top, & (P_k^\perp)^\top &= (P_k^\perp)^2 = P_k^\perp. \end{aligned} \tag{18}$$

Consequently $D_k^\dagger P_k^\perp = 0$ and

$$\boldsymbol{b} = Q_k Q_k^\top \boldsymbol{b} + D_k^\dagger D_k \boldsymbol{b}, \qquad D_k \boldsymbol{b}(\boldsymbol{w}) = (I_{m_k} - P_k^\perp)\boldsymbol{w}, \qquad \widetilde{X}_k P_k^\perp = 0. \tag{19}$$

C.1.1 Profiling and prediction decomposition

For a fixed $\boldsymbol{w}$, the normal equation for the background is

$$(XQ_k)^\top \{XQ_k \boldsymbol{a} - (\boldsymbol{y} - XD_k^\dagger \boldsymbol{w})\} = \boldsymbol{0}.$$

The minimum-norm solution is

$$\boldsymbol{a} = (XQ_k)^\dagger (\boldsymbol{y} - XD_k^\dagger \boldsymbol{w}).$$

Substituting it into (2) gives the reconstructed coefficient vector after profiling, and hence

$$X\boldsymbol{b}(\boldsymbol{w}) = \Pi_k X D_k^\dagger \boldsymbol{w} + (I_n - \Pi_k)\boldsymbol{y},$$

$$X\boldsymbol{b}(\boldsymbol{w}) - \boldsymbol{y} = \Pi_k (X D_k^\dagger \boldsymbol{w} - \boldsymbol{y}), \tag{20}$$

$$\frac{X\{\boldsymbol{b}(\boldsymbol{w}) - \boldsymbol{b}^\star\}}{\sqrt{n}} = \widetilde{X}_k(\boldsymbol{w} - \boldsymbol{w}^\star) + \frac{(I_n - \Pi_k)\boldsymbol{\varepsilon}}{\sqrt{n}}. \tag{21}$$

The last equality uses $\Pi_k X Q_k = 0$ and $\boldsymbol{w}^\star = D_k \boldsymbol{b}^\star$. Since $\Pi_k \widetilde{X}_k = \widetilde{X}_k$ and $\Pi_k(I_n - \Pi_k) = 0$, it also gives the exact identity

$$\mathrm{Pred}_n(\boldsymbol{b}(\boldsymbol{w})) = \|\widetilde{X}_k(\boldsymbol{w} - \boldsymbol{w}^\star)\|_2^2 + \frac{\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2}{n}. \tag{22}$$

These statements hold even when XQ_k is rank deficient, although coefficient directions may then be unidentifiable.

Residual rank and background variance. By (19) and $\Pi_k X Q_k = 0$,

$$\Pi_k X = \Pi_k X D_k^\dagger D_k, \qquad \mathrm{range}(\Pi_k X D_k^\dagger) = \mathrm{range}(\Pi_k X).$$

The reverse range inclusion uses $\Pi_k X D_k^\dagger = (\Pi_k X) D_k^\dagger$. Since $\mathrm{range}(XQ_k) \subseteq \mathrm{range}(X)$, orthogonal projection removes exactly $\mathrm{rank}(XQ_k)$ dimensions from $\mathrm{range}(X)$. Therefore

$$\mathrm{rank}(\widetilde{X}_k) = \mathrm{rank}(\Pi_k X) = \mathrm{rank}(X) - \mathrm{rank}(XQ_k). \tag{23}$$

In particular, full row rank of X and full column rank of XQ_k leave $n - r_k$ residual directions. This is compatible with $p \gg n$.

Assumption 1 implies $\mathbb{E}[\boldsymbol{\varepsilon} \mid X] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid X) \preceq \sigma^2 I_n$. Taking conditional expectations in the background term of (22) gives

$$\mathbb{E}\left[\frac{\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2}{n} \,\middle|\, X\right] = \frac{\mathrm{tr}\{(I_n - \Pi_k)\,\mathrm{Cov}(\boldsymbol{\varepsilon} \mid X)\}}{n} \le \frac{\sigma^2\,\mathrm{rank}(XQ_k)}{n}. \tag{24}$$

Equality holds if $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid X) = \sigma^2 I_n$. The remaining empirical prediction term may have bias and depends on the fitted differences.

C.2 Existence of a backtracking step

Define $f(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{L}_k(\boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2})$. This polynomial has locally Lipschitz gradient, and the chain rule gives

$$\nabla f = \begin{bmatrix} 2\boldsymbol{u} \odot \boldsymbol{g} \\ -2\boldsymbol{v} \odot \boldsymbol{g} \end{bmatrix}, \qquad \|\nabla f\|_2^2 = 4\sum_e (u_e^2 + v_e^2) g_e^2, \qquad \boldsymbol{g} = \nabla \mathcal{L}_k(\boldsymbol{w}). \tag{25}$$

For a finite current point, choose a neighborhood with gradient Lipschitz constant L_{loc}. For a small enough step staying in this neighborhood, the descent lemma gives

$$f((\boldsymbol{u}, \boldsymbol{v}) - \eta \nabla f) \le f(\boldsymbol{u}, \boldsymbol{v}) - \eta(1 - L_{\mathrm{loc}}\eta/2)\|\nabla f\|_2^2.$$

Taking $\eta \le 2(1 - c_A)/L_{\mathrm{loc}}$ ensures Armijo decrease; reducing it further ensures $2\eta\|\boldsymbol{g}\|_\infty < 1$. Both factor multipliers are then positive, so induction preserves strictly positive factors at every finite iterate. If $\boldsymbol{g} = \mathbf{0}$, the update is stationary.

C.3 Finite-step comparison

Let $s_T = \sum_{t<T} \eta_t$ and $\varphi(x) = -\log(1 - x) - x$ for $|x| < 1$. For a comparator $\bar{\boldsymbol{w}} \in \mathbb{R}^{m_k}$, put $\bar{w}_e^+ = \max(\bar{w}_e, 0)$ and $\bar{w}_e^- = \max(-\bar{w}_e, 0)$. Define

$$K_\alpha(\bar{\boldsymbol{w}}) = \sum_e [|\bar{w}_e| \log(|\bar{w}_e|/\alpha^2) - |\bar{w}_e| + 2\alpha^2],$$

$$\mathcal{E}_T(\bar{\boldsymbol{w}}) = \frac{1}{s_T}\left\{\frac{K_\alpha(\bar{\boldsymbol{w}})}{4} + \sum_{t<T}\eta_t^2 \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \right. \tag{26}$$

$$\left. + \frac{1}{2}\sum_{t<T,e} [\bar{w}_e^+ \varphi(2\eta_t g_{t,e}) + \bar{w}_e^- \varphi(-2\eta_t g_{t,e})] \right\}.$$

We use $0 \log 0 = 0$. The last term retains the exact discrete logarithmic remainder.

Lemma C.1 (Finite-step comparison)**.** *For every $\bar{\boldsymbol{w}}$ and finite executed prefix $T \geq 1$ of the balanced, positive-initialization GD updates (3) satisfying (4),*

$$\mathcal{L}_k(\boldsymbol{w}_T) - \mathcal{L}_k(\bar{\boldsymbol{w}}) \leq \mathcal{E}_T(\bar{\boldsymbol{w}}). \tag{27}$$

Proof of Lemma C.1. For a fixed comparator $\bar{\boldsymbol{w}}$, use its positive and negative parts to define

$$\mathcal{D}_t = \sum_e \left[\bar{w}_e^+ \log \frac{\bar{w}_e^+}{u_{t,e}^2} - \bar{w}_e^+ + u_{t,e}^2 + \bar{w}_e^- \log \frac{\bar{w}_e^-}{v_{t,e}^2} - \bar{w}_e^- + v_{t,e}^2 \right]. \tag{28}$$

For $a \geq 0, b > 0$, $a \log(a/b) - a + b \geq 0$, with the continuous convention at $a = 0$. Hence $\mathcal{D}_t \geq 0$ and $\mathcal{D}_0 = K_\alpha(\bar{\boldsymbol{w}})$.

For each coordinate set $x_e = 2\eta_t g_{t,e}$. The factor squares satisfy

$$u_{t+1,e}^2 = u_{t,e}^2(1 - x_e)^2, \qquad v_{t+1,e}^2 = v_{t,e}^2(1 + x_e)^2.$$

Inserting these identities into (28), including both the logarithmic and linear terms, gives

$$\begin{aligned} \mathcal{D}_{t+1} - \mathcal{D}_t = \sum_e \big[& -2\bar{w}_e^+ \log(1 - x_e) - 2\bar{w}_e^- \log(1 + x_e) \\ & - 2x_e w_{t,e} + x_e^2(u_{t,e}^2 + v_{t,e}^2) \big]. \end{aligned} \tag{29}$$

Since $-\log(1 - x) = x + \varphi(x)$ and $-\log(1 + x) = -x + \varphi(-x)$,

$$\begin{aligned} \mathcal{D}_{t+1} - \mathcal{D}_t = & -4\eta_t \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \bar{\boldsymbol{w}} \rangle + 4\eta_t^2 \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \\ & + 2 \sum_e [\bar{w}_e^+ \varphi(2\eta_t g_{t,e}) + \bar{w}_e^- \varphi(-2\eta_t g_{t,e})]. \end{aligned} \tag{30}$$

The convexity step can be seen directly from the quadratic loss:

$$\mathcal{L}_k(\bar{\boldsymbol{w}}) - \mathcal{L}_k(\boldsymbol{w}_t) - \langle \boldsymbol{g}_t, \bar{\boldsymbol{w}} - \boldsymbol{w}_t \rangle = \tfrac{1}{2} \|\widetilde{X}_k(\bar{\boldsymbol{w}} - \boldsymbol{w}_t)\|_2^2 + \tfrac{\gamma}{2} \|P_k^\perp(\bar{\boldsymbol{w}} - \boldsymbol{w}_t)\|_2^2 \geq 0. \tag{31}$$

Sum (30) over $t < T$ and use this inequality and $\mathcal{D}_T \geq 0$:

$$\begin{aligned} \sum_{t<T} \eta_t \{\mathcal{L}_k(\boldsymbol{w}_t) - \mathcal{L}_k(\bar{\boldsymbol{w}})\} \leq & \frac{\mathcal{D}_0 - \mathcal{D}_T}{4} + \sum_{t<T} \eta_t^2 \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \\ & + \frac{1}{2} \sum_{t<T,e} [\bar{w}_e^+ \varphi(2\eta_t g_{t,e}) + \bar{w}_e^- \varphi(-2\eta_t g_{t,e})] \leq s_T \mathcal{E}_T(\bar{\boldsymbol{w}}). \end{aligned} \tag{32}$$

Monotonicity gives $s_T\{\mathcal{L}_k(\boldsymbol{w}_T) - \mathcal{L}_k(\bar{\boldsymbol{w}})\} \leq \sum_{t<T} \eta_t \{\mathcal{L}_k(\boldsymbol{w}_t) - \mathcal{L}_k(\bar{\boldsymbol{w}})\}$. Dividing by $s_T > 0$ proves the claim. Convexity was used in $\boldsymbol{w}$, not in the factors.

Remainders and numerical qualifications. For $|x| < 1$, the power series and nonnegativity give

$$0 \leq \varphi(x) = \sum_{j=2}^{\infty} \frac{x^j}{j} \leq \sum_{j=2}^{\infty} \frac{|x|^j}{2} = \frac{x^2}{2(1 - |x|)}.$$

The logarithmic part of $\mathcal{E}_T(\bar{\boldsymbol{w}})$ is therefore at most

$$\frac{1}{s_T} \sum_{t<T} \frac{\eta_t^2 \|\bar{\boldsymbol{w}}\|_1 \|\boldsymbol{g}_t\|_\infty^2}{1 - 2\eta_t \|\boldsymbol{g}_t\|_\infty}.$$

If $\|\bar{\boldsymbol{w}}\|_1 \leq \vartheta$, then $K_\alpha(\bar{\boldsymbol{w}}) \leq \vartheta \max\{\log(\vartheta/\alpha^2), 0\} + 2m_k\alpha^2$, using $\sum_e |\bar{w}_e| \log |\bar{w}_e| \leq \|\bar{\boldsymbol{w}}\|_1 \log \|\bar{\boldsymbol{w}}\|_1$. The factor product satisfies exactly

$$u_{T,e} v_{T,e} = \alpha^2 \prod_{t<T} (1 - 4\eta_t^2 g_{t,e}^2).$$

Acceptance slack. Suppose the update (3) has positive factors, but the acceptance condition is

$$\mathcal{L}_k(\boldsymbol{w}_{t+1}) \le \mathcal{L}_k(\boldsymbol{w}_t) - 4c_A\eta_t \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 + \mu_t, \qquad \mu_t \ge 0. \tag{33}$$

The KL identity and (32) are unchanged. Instead of exact monotonicity, use $\mathcal{L}_k(\boldsymbol{w}_T) \le \mathcal{L}_k(\boldsymbol{w}_t) + \sum_{j=t}^{T-1} \mu_j$. Thus, defining the following quantity only for this extension,

$$\mathcal{E}_T^{\mathrm{num}}(\bar{\boldsymbol{w}}) = \mathcal{E}_T(\bar{\boldsymbol{w}}) + \frac{1}{s_T} \sum_{t<T} \eta_t \sum_{j=t}^{T-1} \mu_j, \tag{34}$$

we obtain

$$\mathcal{L}_k(\boldsymbol{w}_T) - \mathcal{L}_k(\bar{\boldsymbol{w}}) \le \mathcal{E}_T^{\mathrm{num}}(\bar{\boldsymbol{w}}).$$

The subsequent path-dependent prediction and coefficient-error bounds derived from this comparison remain valid with $\mathcal{E}_T^{\mathrm{num}}$ replacing $\mathcal{E}_T$. This substitution alone does not extend Theorem 2 to arbitrary relaxed trajectories; see Appendix B.6. For the implemented acceptance rule, $\mu_t = 10^{-14}\max\{1, \mathcal{L}_k(\boldsymbol{w}_t)\}$. This is an exact-arithmetic statement about a test with slack.

Corollary C.2 (Energy and factor-product control)**.** *For a positive-factor trajectory satisfying* (33) *with* $0 < \eta_t \le \eta_{\max}$,

$$\sum_{t<T} \eta_t^2 \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \le \frac{\eta_{\max}}{4c_A} \left\{ \mathcal{L}_k(\boldsymbol{w}_0) - \mathcal{L}_k(\boldsymbol{w}_T) + \sum_{t<T} \mu_t \right\}, \tag{35}$$

$$\begin{aligned} \mathcal{E}_T^{\mathrm{num}}(\bar{\boldsymbol{w}}) &\le \frac{K_\alpha(\bar{\boldsymbol{w}})}{4s_T} + \frac{\eta_{\max}\{\mathcal{L}_k(\boldsymbol{w}_0) + \sum_{t<T}\mu_t\}}{4c_A s_T} \\ &\quad + \frac{1}{2s_T} \sum_e |\bar{w}_e| \log \frac{\alpha^2}{u_{T,e} v_{T,e}} + \frac{1}{s_T} \sum_{t<T} \eta_t \sum_{j=t}^{T-1} \mu_j. \end{aligned} \tag{36}$$

For exact Armijo descent all μ_t terms vanish.

Proof of Corollary C.2. Rearranging (33) and summing gives

$$4c_A \sum_{t<T} \eta_t \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \le \mathcal{L}_k(\boldsymbol{w}_0) - \mathcal{L}_k(\boldsymbol{w}_T) + \sum_{t<T} \mu_t.$$

Use $\eta_t^2 \le \eta_{\max}\eta_t$ to obtain (35). Next, for $|x| < 1$,

$$\varphi(x) + \varphi(-x) = -\log(1 - x^2), \qquad \varphi(x) \ge 0, \quad \varphi(-x) \ge 0.$$

Since $\bar{w}_e^+ + \bar{w}_e^- = |\bar{w}_e|$, the logarithmic term satisfies

$$\begin{aligned} &\sum_{t<T,e} \{\bar{w}_e^+ \varphi(2\eta_t g_{t,e}) + \bar{w}_e^- \varphi(-2\eta_t g_{t,e})\} \\ &\quad \le \sum_e |\bar{w}_e| \sum_{t<T} -\log(1 - 4\eta_t^2 g_{t,e}^2) = \sum_e |\bar{w}_e| \log \frac{\alpha^2}{u_{T,e} v_{T,e}}. \end{aligned}$$

Insert this and (35) into (34), and use $\mathcal{L}_k(\boldsymbol{w}_T) \ge 0$. □

This corollary controls the accumulated quadratic term by the loss budget and isolates the product drift. In particular, the bound can be conservative for a small Armijo constant.

C.4 Empirical prediction and consistency

Locally write $\boldsymbol{h}_T = \boldsymbol{b}(\boldsymbol{w}_T) - \boldsymbol{b}^\star$ and let $\boldsymbol{e}_j$ be the jth standard basis vector in $\mathbb{R}^{m_k}$. Set

$$\nu_\delta = \sigma \max_j \|\widetilde{X}_k \boldsymbol{e}_j\|_2 \sqrt{\frac{2\log(4m_k/\delta)}{n}}. \tag{37}$$

Lemma C.3 (Empirical prediction). *Under Assumption 1, conditional on X, with probability at least $1-\delta$, simultaneously for all finite executed prefixes $T \geq 1$ of the balanced, positive-initialization GD updates (3) satisfying the positivity and exact Armijo conditions (4),*

$$\|\widetilde{X}_k(\boldsymbol{w}_T - \boldsymbol{w}^\star)\|_2^2 + \gamma\|(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w}_T\|_2^2 \leq 2\nu_\delta\|D_k\boldsymbol{h}_T\|_1 + 2\mathcal{E}_T(\boldsymbol{w}^\star), \tag{38}$$

$$\mathrm{Pred}_n(\boldsymbol{b}(\boldsymbol{w}_T)) \leq 2\nu_\delta\|D_k\boldsymbol{h}_T\|_1 + 2\mathcal{E}_T(\boldsymbol{w}^\star) + \frac{8\sigma^2}{n}\{r_k\log 5 + \log(4/\delta)\}. \tag{39}$$

For prefixes satisfying the relaxed acceptance rule (33), replace $\mathcal{E}_T$ in both bounds by $\mathcal{E}_T^{\mathrm{num}}$ from (34). Moreover, every iterate satisfies the deterministic consistency bound

$$|\|D_k\boldsymbol{b}(\boldsymbol{w}_T)\|_1 - \|\boldsymbol{w}_T\|_1| \leq \sqrt{m_k}\|(I_{m_k} - D_k D_k^\dagger)\boldsymbol{w}_T\|_2. \tag{40}$$

Proof of Lemma C.3. Within this proof put $\boldsymbol{\xi} = \Pi_k\boldsymbol{\varepsilon}/\sqrt{n}$. Profiling gives $\widetilde{\boldsymbol{y}}_k = \widetilde{X}_k\boldsymbol{w}^\star + \boldsymbol{\xi}$ and $P_k^\perp\boldsymbol{w}^\star = \mathbf{0}$. Expanding the loss difference at the truth yields

$$\begin{aligned}\mathcal{L}_k(\boldsymbol{w}_T) - \mathcal{L}_k(\boldsymbol{w}^\star) &= \tfrac{1}{2}\|\widetilde{X}_k(\boldsymbol{w}_T - \boldsymbol{w}^\star) - \boldsymbol{\xi}\|_2^2 - \tfrac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \tfrac{\gamma}{2}\|P_k^\perp\boldsymbol{w}_T\|_2^2 \\ &= \tfrac{1}{2}\|\widetilde{X}_k(\boldsymbol{w}_T - \boldsymbol{w}^\star)\|_2^2 - \langle\widetilde{X}_k^\top\boldsymbol{\xi}, \boldsymbol{w}_T - \boldsymbol{w}^\star\rangle + \tfrac{\gamma}{2}\|P_k^\perp\boldsymbol{w}_T\|_2^2. \end{aligned}\tag{41}$$

The comparison lemma consequently implies

$$\tfrac{1}{2}\|\widetilde{X}_k(\boldsymbol{w}_T - \boldsymbol{w}^\star)\|_2^2 + \tfrac{\gamma}{2}\|P_k^\perp\boldsymbol{w}_T\|_2^2 \leq \langle\widetilde{X}_k^\top\boldsymbol{\xi}, \boldsymbol{w}_T - \boldsymbol{w}^\star\rangle + \mathcal{E}_T(\boldsymbol{w}^\star). \tag{42}$$

For each coordinate j, $\Pi_k\widetilde{X}_k = \widetilde{X}_k$ and the joint noise assumption imply

$$\mathbb{E}\Big[\exp\{t(\widetilde{X}_k\boldsymbol{e}_j)^\top\boldsymbol{\xi}\} \mid X\Big] \leq \exp\left(\frac{t^2\sigma^2\|\widetilde{X}_k\boldsymbol{e}_j\|_2^2}{2n}\right).$$

The two-sided Chernoff bound and a union bound over m_k coordinates give

$$\mathbb{P}(\|\widetilde{X}_k^\top\boldsymbol{\xi}\|_\infty > \nu_\delta \mid X) \leq 2m_k\exp\{-\log(4m_k/\delta)\} = \delta/2.$$

Zero columns are identically zero. Independence of the projected noise coordinates is unnecessary. On this single event, using $\widetilde{X}_k P_k^\perp = 0$,

$$\begin{aligned}\langle\widetilde{X}_k^\top\boldsymbol{\xi}, \boldsymbol{w}_T - \boldsymbol{w}^\star\rangle &= \langle\widetilde{X}_k^\top\boldsymbol{\xi}, (I_{m_k} - P_k^\perp)\boldsymbol{w}_T - \boldsymbol{w}^\star\rangle \\ &= \langle\widetilde{X}_k^\top\boldsymbol{\xi}, D_k\boldsymbol{b}(\boldsymbol{w}_T) - D_k\boldsymbol{b}^\star\rangle \leq \nu_\delta\|D_k\boldsymbol{h}_T\|_1. \end{aligned}\tag{43}$$

Substitution into (42) proves (38).

For the background, a $1/2$-net $\mathcal{N}$ of the unit sphere in $\mathrm{range}(I_n - \Pi_k)$ has size at most 5^{r_k}. Approximating the normalized projected noise by a net point gives

$$\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2 \leq 2\max_{\boldsymbol{\theta}\in\mathcal{N}}|\boldsymbol{\theta}^\top\boldsymbol{\varepsilon}|.$$

With $t_0 = r_k\log 5 + \log(4/\delta)$,

$$\mathbb{P}\{\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2 > 8\sigma^2 t_0 \mid X\} \leq 2\,5^{r_k}e^{-t_0} = \delta/2.$$

The zero-rank case is immediate. Combining both events with (22) proves (39). Neither event depends on T, so no union bound over the trajectory is required.

Finally, the reverse triangle inequality gives

$$|\|D_k\boldsymbol{b}(\boldsymbol{w}_T)\|_1 - \|\boldsymbol{w}_T\|_1| \leq \|P_k^\perp\boldsymbol{w}_T\|_1 \leq \sqrt{m_k}\|P_k^\perp\boldsymbol{w}_T\|_2,$$

which proves (40). At the initial background, $s_0 = 0$ and $\mathcal{E}_0$ is not used; instead,

$$\mathrm{Pred}_n(\boldsymbol{b}(\mathbf{0})) = \|\widetilde{X}_k\boldsymbol{w}^\star\|_2^2 + \|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2/n. \tag{44}$$

C.5 Proof of Theorem 1: curvature and risk

Lemma C.4 (Curvature for every difference-space representation)**.** *Under Assumption 2, if* $n \geq c_0\omega^4\{r_k + \log(8m_k/\delta)\}$*, then with probability at least* $1-\delta$*, simultaneously for every* $\boldsymbol{a} \in \mathbb{R}^{r_k}$ *and* $\boldsymbol{z} \in \mathbb{R}^{m_k}$*, the vector* $\boldsymbol{h} = Q_k\boldsymbol{a} + D_k^\dagger \boldsymbol{z}$ *satisfies*

$$\frac{\|X\boldsymbol{h}\|_2^2}{n} \geq \tfrac{1}{4}\boldsymbol{h}^\top\Sigma\boldsymbol{h} - \frac{c_1\omega^4\ell_k^2\log(2m_k)}{n}\|\boldsymbol{z}\|_1^2. \tag{45}$$

This event makes XQ_k *full column rank. Taking* $\boldsymbol{z} = D_k\boldsymbol{h}$ *also yields*

$$\frac{\|X\boldsymbol{h}\|_2^2}{n} \geq \tfrac{1}{4}\boldsymbol{h}^\top\Sigma\boldsymbol{h} - \frac{c_1\omega^4\ell_k^2\log(2m_k)}{n}\|D_k\boldsymbol{h}\|_1^2 \quad \textit{for all } \boldsymbol{h} \in \mathbb{R}^p. \tag{46}$$

Proof of Lemma C.4. For this proof, denote the two coefficients in (46) by

$$\kappa = \tfrac{1}{4}, \qquad \tau = c_1\omega^4\ell_k^2\log(2m_k)/n. \tag{47}$$

The claimed inequality then reads $\|X\boldsymbol{h}\|_2^2/n \geq \kappa\boldsymbol{h}^\top\Sigma\boldsymbol{h} - \tau\|D_k\boldsymbol{h}\|_1^2$.

Step 1: normalized difference geometry. The whitened rows of $X\Sigma^{-1/2}$ are independent isotropic sub-Gaussian vectors with sub-Gaussian norm at most $C_0\omega$. For a deterministic subset $\mathcal{T}$ of the unit sphere, Theorem 1.4 and Corollary 1.5 of Liaw et al. (2017) imply, with probability at least $1-\delta$,

$$\sup_{\boldsymbol{t}\in\mathcal{T}}\left|\frac{\|X\Sigma^{-1/2}\boldsymbol{t}\|_2}{\sqrt{n}} - 1\right| \leq \frac{C_{\mathrm{md}}\omega^2}{\sqrt{n}}\left(\mathbb{E}\sup_{\boldsymbol{t}\in\mathcal{T}}|\boldsymbol{\zeta}^\top\boldsymbol{t}| + \sqrt{\log(4/\delta)}\right), \tag{48}$$

where $\boldsymbol{\zeta} \sim N(\boldsymbol{0}, I_p)$ is auxiliary Gaussian randomness. The absolute constant C_{md}, used only in this proof, includes the conversion from the mgf condition to the sub-Gaussian norm. For a deterministic $a > 0$, take

$$\mathcal{T} = \{\Sigma^{1/2}(Q_k\boldsymbol{a} + D_k^\dagger\boldsymbol{z}) : \|\Sigma^{1/2}(Q_k\boldsymbol{a} + D_k^\dagger\boldsymbol{z})\|_2 = 1,\ \|\boldsymbol{z}\|_1 \leq a\}.$$

Let P be the orthogonal projection onto $\mathrm{range}(\Sigma^{1/2}Q_k)$, of rank r_k. By its defining representation, every $\boldsymbol{t} = \Sigma^{1/2}\boldsymbol{h} \in \mathcal{T}$ satisfies

$$\boldsymbol{t} = P\boldsymbol{t} + (I_p - P)\Sigma^{1/2}D_k^\dagger\boldsymbol{z}.$$

Therefore

$$|\boldsymbol{\zeta}^\top\boldsymbol{t}| \leq \|P\boldsymbol{\zeta}\|_2 + \|\boldsymbol{z}\|_1\max_j|\boldsymbol{\zeta}^\top(I_p - P)\Sigma^{1/2}D_k^\dagger\boldsymbol{e}_j|,$$
$$\mathbb{E}\sup_{\boldsymbol{t}\in\mathcal{T}}|\boldsymbol{\zeta}^\top\boldsymbol{t}| \leq \sqrt{r_k} + a\ell_k\sqrt{2\log(2m_k)}, \tag{49}$$

where $\mathbb{E}\|P\boldsymbol{\zeta}\|_2 \leq \sqrt{\mathrm{tr}\,P} = \sqrt{r_k}$ bounds the first term. Each Gaussian contrast in the maximum has variance at most ℓ_k^2; its exponential moment bound gives the displayed maximum bound even when the contrasts are dependent.

Step 2: curvature for all directions. Choose

$$a = \frac{\sqrt{n}}{4\sqrt{2}\,C_{\mathrm{md}}\omega^2\ell_k\sqrt{\log(2m_k)}}.$$

The contribution of $a\ell_k\sqrt{2\log(2m_k)}$ to (48) is $1/4$. If $n \geq 32C_{\mathrm{md}}^2\omega^4\{r_k+\log(4/\delta)\}$, the remaining contribution is at most $1/4$, since $\sqrt{r_k} + \sqrt{\log(4/\delta)} \leq \sqrt{2\{r_k + \log(4/\delta)\}}$. Thus the lemma's sample condition suffices with $c_0 \geq 32C_{\mathrm{md}}^2$. For any representation $\boldsymbol{h} = Q_k\boldsymbol{a} + D_k^\dagger\boldsymbol{z} \neq \boldsymbol{0}$ with $\|\boldsymbol{z}\|_1 \leq a\|\Sigma^{1/2}\boldsymbol{h}\|_2$, normalize by $\|\Sigma^{1/2}\boldsymbol{h}\|_2$ to obtain

$$\frac{\|X\boldsymbol{h}\|_2}{\sqrt{n}\|\Sigma^{1/2}\boldsymbol{h}\|_2} \geq \frac{1}{2}, \qquad \frac{\|X\boldsymbol{h}\|_2^2}{n} \geq \frac{1}{4}\boldsymbol{h}^\top\Sigma\boldsymbol{h}.$$

For every remaining direction,

$$\frac{1}{4}\boldsymbol{h}^\top \Sigma \boldsymbol{h} - \frac{\|\boldsymbol{z}\|_1^2}{4a^2} < 0 \le \frac{\|X\boldsymbol{h}\|_2^2}{n}.$$

Both cases, and $\boldsymbol{h} = \mathbf{0}$, therefore prove (45) with $c_1 = 8C_{\mathrm{md}}^2$. In the local notation (47), this gives $\tau = 1/(4a^2) = 8C_{\mathrm{md}}^2\omega^4\ell_k^2\log(2m_k)/n$. The confidence term is absorbed by the sample-size requirement, so the tolerance itself needs no additional $\log(1/\delta)$ factor. Setting $\boldsymbol{z} = \mathbf{0}$ proves injectivity on $\mathrm{range}(Q_k)$ because $\Sigma \succ 0$; setting $\boldsymbol{z} = D_k\boldsymbol{h}$ and $\boldsymbol{a} = Q_k^\top \boldsymbol{h}$ proves (46).

Step 3: the noise scale and population error. For $Z_{i,j} = \boldsymbol{x}_i^\top D_k^\dagger \boldsymbol{e}_j$, write $\ell_{k,j} = \|\Sigma^{1/2} D_k^\dagger \boldsymbol{e}_j\|_2$. Then $\mathbb{E}Z_{i,j}^2 = \ell_{k,j}^2$ and the centered square has sub-exponential norm at most $C_0\omega^2\ell_{k,j}^2$. Bernstein's inequality yields, for $t > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_i Z_{i,j}^2 > \ell_{k,j}^2\left[1 + C_0\omega^2\left(\sqrt{t/n} + t/n\right)\right]\right\} \le 2e^{-t}.$$

Zero-variance columns vanish almost surely. Use $t = \log(8m_k/\delta)$, take a union bound, and use contraction by Π_k:

$$\max_j \|\widetilde{X}_k \boldsymbol{e}_j\|_2 \le \max_j \frac{\|X D_k^\dagger \boldsymbol{e}_j\|_2}{\sqrt{n}} \le C_0\ell_k. \tag{50}$$

The last inequality holds with failure probability at most $\delta/4$ when $n \gtrsim \omega^4 \log(8m_k/\delta)$.

For any design satisfying (46), the coefficient error $\boldsymbol{h}_T = \boldsymbol{b}(\boldsymbol{w}_T) - \boldsymbol{b}^\star$ obeys

$$\begin{aligned}\kappa\mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)) &\le \mathrm{Pred}_n(\boldsymbol{b}(\boldsymbol{w}_T)) + \tau\|D_k\boldsymbol{h}_T\|_1^2 \\ &\le 2\nu_{\delta/2}\|D_k\boldsymbol{h}_T\|_1 + 2\mathcal{E}_T(\boldsymbol{w}^\star) + \tau\|D_k\boldsymbol{h}_T\|_1^2 + \frac{8\sigma^2}{n}\{r_k\log 5 + \log(8/\delta)\}. \end{aligned}\tag{51}$$

Allocate $\delta/4$ to the curvature event and $\delta/4$ to column concentration. Apply Lemma C.3 at level $\delta/2$, allocating $\delta/4$ each to its coordinate-noise and background-noise events. A union bound over these four events gives total failure probability at most δ. Equations (50) and (51) give the bound in Theorem 1, after adjusting absolute constants. All four events cover every finite iterate. Also,

$$\lambda_{\min}(\Sigma)\|\mathcal{B}(\boldsymbol{w}_T) - \mathcal{B}^\star\|_F^2 = \lambda_{\min}(\Sigma)\|\boldsymbol{h}_T\|_2^2 \le \boldsymbol{h}_T^\top\Sigma\boldsymbol{h}_T = \mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)).$$

The slack-acceptance extension follows by replacing $\mathcal{E}_T$ with (34).

Why the tolerance remains. If $p > n$ and $\Sigma \succ 0$, there is a nonzero $\boldsymbol{h} \in \ker X$ with $\boldsymbol{h}^\top\Sigma\boldsymbol{h} > 0$. Hence a bound $\|X\boldsymbol{h}\|_2^2/n \ge \kappa\boldsymbol{h}^\top\Sigma\boldsymbol{h}$ with $\kappa > 0$ cannot hold for all directions. A cone condition alone does not remove the tolerance: one would additionally need a quantitative compatibility bound, for example $\|D_k\boldsymbol{h}\|_1^2 \le C\,\boldsymbol{h}^\top\Sigma\boldsymbol{h}$ on the relevant error set, and $\tau C < \kappa$ to absorb it. We do not assume that the GD trajectory satisfies such a condition.

C.6 Proof of Theorem 2: explicit steps and budget

All auxiliary symbols in this subsection are local. Fix k, and write $A = \widetilde{X}_k$, $P = P_k^\perp$, $m = m_k$, $r = r_k$, $\boldsymbol{\xi} = \Pi_k\boldsymbol{\varepsilon}/\sqrt{n}$, and $\vartheta = \|\boldsymbol{w}^\star\|_1 \le V$. Thus

$$\widetilde{\boldsymbol{y}}_k = A\boldsymbol{w}^\star + \boldsymbol{\xi}, \qquad P\boldsymbol{w}^\star = \mathbf{0}, \qquad AP = 0. \tag{52}$$

The vector error always admits a representation $\boldsymbol{b}(\boldsymbol{w}) - \boldsymbol{b}^\star = Q_k\boldsymbol{a} + D_k^\dagger(\boldsymbol{w} - \boldsymbol{w}^\star)$ for some possibly data-dependent $\boldsymbol{a}$. Lemma C.4 is uniform over $\boldsymbol{a}$ and the representation vector, so it applies to this error.

Explicit parameters and window. For $0 < \delta < 1/4$, define

$$\rho = 2\sqrt{2}\,\sigma\ell_k\sqrt{\frac{\log(20m/\delta)}{n}} + \frac{\omega^4\ell_k^2 W_\alpha\log(2m)}{n}, \tag{53}$$

$$\beta = 4m\ell_k^2, \qquad \Lambda = 4\ell_k^2V^2 + 8\sigma^2\left\{\log 5 + \frac{\log(20/\delta)}{n}\right\}, \qquad G = \sqrt{2\beta\Lambda}, \tag{54}$$

where ρ is a proof-local stopping scale, not the consistency weight γ. Since $\alpha > 0$, $D_k \neq 0$, and $\Sigma \succ 0$, one has $W_\alpha, \ell_k, \rho > 0$. Use the same value for the initial step and its cap:

$$\eta = \min\left\{\frac{1}{4G}, \frac{\rho}{2G^2}, \frac{1-c_A}{48\beta W_\alpha + G}, \frac{1}{64\rho}\right\}. \tag{55}$$

Terms with zero denominator in the first two entries are interpreted as $+\infty$. The step is positive and finite. Retain the usual backtracking and capped 1.02 trial increase. We prove that the first trial is always accepted up to

$$T_- = \left\lceil \frac{1}{32\rho\eta} \right\rceil, \qquad T_+ = \left\lfloor \frac{1}{16\rho\eta} \right\rfloor. \tag{56}$$

For $A \neq 0$, choose the external iteration budget $T_{\max} \geq T_+$ so that all these updates are scheduled. In particular $T_- \leq T_+$, and the iteration count is explicitly bounded by

$$T_+ \leq \max\left\{\frac{G}{4\rho}, \frac{G^2}{8\rho^2}, \frac{48\beta W_\alpha + G}{16(1-c_A)\rho}, 4\right\}. \tag{57}$$

These are conservative statistical parameters, not the experimental step cap.

Step 1: a deterministic gradient envelope. Initially work on the event

$$\max_j \|A\boldsymbol{e}_j\|_2 \leq 2\ell_k, \qquad \|A^\top \boldsymbol{\xi}\|_\infty \leq \rho, \qquad \mathcal{L}_k(\mathbf{0}) \leq \Lambda. \tag{58}$$

Its probability, together with the necessary curvature and background events, is established below. Orthogonality $AP = 0$ and $\gamma = \|A\|_{\mathrm{op}}^2$ imply

$$\|A^\top A + \gamma P\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}^2 \leq \|A\|_F^2 \leq \beta.$$

Consequently, for every $\boldsymbol{w}$,

$$\begin{aligned}\|\nabla\mathcal{L}_k(\boldsymbol{w})\|_\infty^2 &\leq \|\nabla\mathcal{L}_k(\boldsymbol{w})\|_2^2 \\ &\leq \|[A^\top \sqrt{\gamma}P]\|_{\mathrm{op}}^2\{\|A\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k\|_2^2 + \gamma\|P\boldsymbol{w}\|_2^2\} \leq 2\beta\mathcal{L}_k(\boldsymbol{w}).\end{aligned} \tag{59}$$

Thus $\mathcal{L}_k(\boldsymbol{w}_t) \leq \Lambda$ implies $\|\boldsymbol{g}_t\|_\infty \leq G$.

Step 2: factor mass from generalized KL. Let $\mathcal{D}_t$ be (28) with comparator $\boldsymbol{w}^\star$, and put $M_t = \sum_e (u_{t,e}^2 + v_{t,e}^2)$ only in this proof. For $a \geq 0, b > 0$,

$$a\log(a/b) - a + b \geq b/2 - a\log 2.$$

For $a > 0$, after division by a the difference is $-\log(b/a) - 1 + b/(2a) + \log 2$, minimized at $b/a = 2$; the case $a = 0$ is immediate. Summation therefore gives

$$M_t \leq 2\mathcal{D}_t + 2\vartheta\log 2, \qquad M_t + \vartheta \leq 2\mathcal{D}_t + 3\vartheta. \tag{60}$$

Since $\sum_e |w_e^\star|\log|w_e^\star| \leq \vartheta\log\vartheta$,

$$\mathcal{D}_0 + \tfrac{3}{2}\vartheta \leq W_\alpha, \qquad \vartheta \leq V \leq W_\alpha/2. \tag{61}$$

Indeed, $\mathcal{D}_0 \leq \vartheta\log(\vartheta/\alpha^2) - \vartheta + 2m\alpha^2$ after combining coordinates; using $\vartheta \leq V$ and discarding negative contributions yields (61). The zero-mass case follows by continuity.

Step 3: current mass control implies first-trial Armijo acceptance. Suppose $M_t \leq 6W_\alpha$ and $\mathcal{L}_k(\boldsymbol{w}_t) \leq \Lambda$. Equation (55) gives $2\eta\|\boldsymbol{g}_t\|_\infty \leq 1/2$. On the factor-space segment $(\boldsymbol{u}(s), \boldsymbol{v}(s)) = (\boldsymbol{u}_t, \boldsymbol{v}_t) - s\nabla f(\boldsymbol{u}_t, \boldsymbol{v}_t)$, $0 \leq s \leq \eta$, all coordinate multipliers lie in $[1/2, 3/2]$. Thus its mass satisfies $M(s) \leq 9M_t/4$ and $\|\boldsymbol{w}(s)\|_2 \leq M(s)$. Using $\|\nabla\mathcal{L}_k(\mathbf{0})\|_\infty \leq G$, we obtain

$$\|\nabla\mathcal{L}_k(\boldsymbol{w}(s))\|_\infty \leq \tfrac{9}{4}\beta M_t + G.$$

The exact factor Hessian is

$$\nabla^2 f = J^\top(A^\top A + \gamma P)J + \begin{bmatrix} 2\,\mathrm{diag}(\boldsymbol{g}) & 0 \\ 0 & -2\,\mathrm{diag}(\boldsymbol{g}) \end{bmatrix}, \qquad J = [2\,\mathrm{diag}(\boldsymbol{u}), -2\,\mathrm{diag}(\boldsymbol{v})].$$

Since $\|J\|_{\mathrm{op}}^2 \leq 4M(s)$, throughout this segment

$$\|\nabla^2 f\|_{\mathrm{op}} \leq \tfrac{27}{2}\beta M_t + 2G \leq 81\beta W_\alpha + 2G. \tag{62}$$

Taylor's integral formula and (55) yield

$$f_{t+1} \leq f_t - \eta\left\{1 - \tfrac{\eta}{2}(81\beta W_\alpha + 2G)\right\}\|\nabla f_t\|_2^2 \leq f_t - c_A\eta\|\nabla f_t\|_2^2.$$

Indeed $(81\beta W_\alpha + 2G)/2 \leq 48\beta W_\alpha + G$. The first trial preserves positive factors and satisfies exact Armijo.

Step 4: the KL recursion closes the induction. By (52),

$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle &= \|A(\boldsymbol{w}_t - \boldsymbol{w}^\star)\|_2^2 + \gamma\|P\boldsymbol{w}_t\|_2^2 - \langle A^\top \boldsymbol{\xi}, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle \\ &\geq -\rho(M_t + \vartheta). \end{aligned} \tag{63}$$

This uses $\|\boldsymbol{w}_t\|_1 \leq M_t$, not contraction by $I_m - P$. For $|x| \leq 1/2$, the exact remainder satisfies

$$0 \leq \varphi(x) = -\log(1-x) - x \leq \frac{x^2}{2(1-|x|)} \leq x^2.$$

Inserting this and (63) into (30) gives

$$\begin{aligned}\mathcal{D}_{t+1} &\leq \mathcal{D}_t + 4\eta\rho(M_t + \vartheta) + 4\eta^2 G^2 M_t + 8\eta^2 G^2 \vartheta \\ &\leq \mathcal{D}_t + 8\eta\rho(M_t + \vartheta), \end{aligned} \tag{64}$$

where $\eta G^2 \leq \rho/2$ was used. Set $Z_t = \mathcal{D}_t + 3\vartheta/2$. Equations (60)–(61) imply

$$Z_{t+1} \leq (1 + 16\eta\rho) Z_t, \qquad Z_t \leq W_\alpha e^{16\rho t\eta}.$$

For $t\eta \leq 1/(16\rho)$,

$$M_t \leq 2Z_t \leq 2eW_\alpha < 6W_\alpha. \tag{65}$$

This proves a joint induction: the initial mass is in the envelope; the current envelope ensures acceptance; acceptance preserves the loss envelope; the KL recursion bounds the next mass. All steps through T_+ are therefore executed with $\eta_t = \eta$. The capped trial increase leaves the next first trial equal to η. This proves finite iteration attainability without assuming a lower bound on accepted steps.

Step 5: product drift and the last-iterate comparison. The discrete product identity and $4\eta^2 g_{t,e}^2 \leq 1/4$ give

$$\begin{aligned}0 \leq \log \frac{\alpha^2}{u_{T,e} v_{T,e}} &= \sum_{t<T} -\log(1 - 4\eta^2 g_{t,e}^2) \\ &\leq \tfrac{16}{3}\eta^2 G^2 T \leq \tfrac{8}{3}\rho T\eta \leq \tfrac{1}{6}, \qquad T \leq T_+. \end{aligned} \tag{66}$$

In particular, $u_{T,e} v_{T,e} \geq e^{-1/6}\alpha^2$. For the comparison remainder (26), write $s_T = T\eta$ locally. Its energy and logarithmic terms obey

$$\frac{1}{s_T}\sum_{t<T}\eta^2 \sum_e (u_{t,e}^2 + v_{t,e}^2) g_{t,e}^2 \leq 6W_\alpha \eta G^2 \leq 3W_\alpha \rho,$$

$$\frac{1}{2s_T}\sum_{t<T,e}\left[w_e^{\star+}\varphi(2\eta g_{t,e}) + w_e^{\star-}\varphi(-2\eta g_{t,e})\right] \leq 2\vartheta\eta G^2 \leq \vartheta\rho.$$

For $T \geq T_-$, $K_\alpha(\boldsymbol{w}^\star)/(4s_T) \leq 8W_\alpha\rho$. Hence throughout the window,

$$\mathcal{E}_T(\boldsymbol{w}^\star) \leq 12W_\alpha\rho, \qquad \|\boldsymbol{w}_T - \boldsymbol{w}^\star\|_1 \leq M_T + \vartheta \leq \tfrac{13}{2}W_\alpha. \tag{67}$$

The passage to the last iterate uses exact loss monotonicity in Lemma C.1; it does not use a continuous-flow approximation. Expanding the quadratic loss as in (42) now yields

$$\|A(\boldsymbol{w}_T - \boldsymbol{w}^\star)\|_2^2 + \gamma\|P\boldsymbol{w}_T\|_2^2 \leq 2\rho\|\boldsymbol{w}_T - \boldsymbol{w}^\star\|_1 + 2\mathcal{E}_T(\boldsymbol{w}^\star) \leq 37W_\alpha\rho. \tag{68}$$

Step 6: a common high-probability event. For clarity, we give concentration details with numerical slack in the constants. If a mean-zero scalar Z has variance s^2 and the design mgf bound with scale ωs, then

$$\mathbb{P}(|Z| > u) \leq 2e^{-u^2/(2\omega^2 s^2)}, \qquad \mathbb{E}|Z|^{2a} \leq 2(2\omega^2 s^2)^a a! \quad (a = 2, 3, \ldots).$$

For $Y = Z^2 - s^2$, this implies $\mathbb{E}|Y|^a \leq a! b^a$ with $b = 8\omega^2 s^2$, and $\mathbb{E}e^{tY} \leq \exp\{t^2 b^2/(1-|t|b)\}$ when $|t|b < 1$. Multiplying mgfs over independent rows and taking $t = \min\{x/(8b^2), 1/(4b)\}$ for each tail gives

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n Z_i^2 - s^2\right| > x\right) \leq 2\exp\left[-\frac{n}{16}\min\left\{\frac{x^2}{b^2}, \frac{x}{b}\right\}\right]. \tag{69}$$

Zero-variance contrasts vanish almost surely. Apply this to $Z_i = \boldsymbol{x}_i^\top D_k^\dagger \boldsymbol{e}_j$, whose variance is at most ℓ_k^2, with x equal to three times its variance. If $n \geq 1024\omega^4 \log(20m/\delta)$, a union bound and projection contraction give, with failure probability at most $\delta/10$,

$$\max_j \|A\boldsymbol{e}_j\|_2 \leq \max_j \|XD_k^\dagger \boldsymbol{e}_j\|_2/\sqrt{n} \leq 2\ell_k.$$

For joint conditional sub-Gaussian noise, a $1/2$-net of the unit sphere in a fixed a-dimensional subspace has at most 5^a points. Conditional on X, the linear-contrast tail bound and a union bound therefore give

$$\|\boldsymbol{\varepsilon}\|_2^2/n \leq 8\sigma^2\{\log 5 + \log(20/\delta)/n\}, \tag{70}$$

$$\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2/n \leq \frac{8\sigma^2}{n}\{r\log 5 + \log(20/\delta)\}, \tag{71}$$

$$\|A^\top \boldsymbol{\xi}\|_\infty \leq \sigma \max_j \|A\boldsymbol{e}_j\|_2 \sqrt{2\log(20m/\delta)/n}. \tag{72}$$

Each failure probability is at most $\delta/10$. In the last display, the conditional noise proxy for coordinate j is $\sigma\|A\boldsymbol{e}_j\|_2/\sqrt{n}$. On these events,

$$\mathcal{L}_k(\mathbf{0}) = \tfrac{1}{2}\|A\boldsymbol{w}^\star + \boldsymbol{\xi}\|_2^2 \leq \|A\boldsymbol{w}^\star\|_2^2 + \|\boldsymbol{\xi}\|_2^2 \leq 4\ell_k^2 V^2 + \|\boldsymbol{\varepsilon}\|_2^2/n \leq \Lambda.$$

Thus (58) holds. Apply Lemma C.4 with failure probability $\delta/10$, enlarging the absolute sample constant to replace its logarithm by $\log(20m/\delta)$. The resulting five events fail with total probability at most $\delta/2$. Full background rank follows from (45) with representation vector zero.

Step 7: population prediction risk and the degenerate branch. Let $\tau = c_1\omega^4\ell_k^2 \log(2m)/n$ be the tolerance coefficient in (45), as in (47). Using the error representation preceding (53),

$$\tfrac{1}{4}\mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)) \leq \mathrm{Pred}_n(\boldsymbol{b}(\boldsymbol{w}_T)) + \tau\|\boldsymbol{w}_T - \boldsymbol{w}^\star\|_1^2.$$

Equations (22), (68), and (71) first give the empirical prediction bound

$$\mathrm{Pred}_n(\boldsymbol{b}(\boldsymbol{w}_T)) \leq 37W_\alpha\rho + \frac{8\sigma^2}{n}\{r\log 5 + \log(20/\delta)\}. \tag{73}$$

Then (67) yields

$$\mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T)) \leq 148W_\alpha\rho + 169\tau W_\alpha^2 + \frac{32\sigma^2}{n}\{r\log 5 + \log(20/\delta)\}. \tag{74}$$

Substituting (53) proves (5). All remaining constants are absolute; the ω^4 dependence is explicit. Finally, $\lambda_{\min}(\Sigma)\|\mathcal{B}(\boldsymbol{w}_T) - \mathcal{B}^\star\|_F^2 \leq \mathcal{R}(\boldsymbol{b}(\boldsymbol{w}_T))$.

If $A = 0$, the prescribed branch returns the background. Its empirical error is the background-noise term alone, and its error representation uses $\boldsymbol{z} = -\boldsymbol{w}^\star$. Thus

$$\mathcal{R}(\boldsymbol{b}(\mathbf{0})) \leq 4\|(I_n - \Pi_k)\boldsymbol{\varepsilon}\|_2^2/n + 4\tau V^2,$$

which is covered by (5).

C.7 Checkpoint coverage for Theorem 2 and validation selection

For the nondegenerate run, put $a = 1/(32\rho\eta)$ locally. The last entry of (55) gives $a \geq 2$. Taking $j = \lceil \log_2 a \rceil$ yields $a \leq 2^j \leq 2a$, and hence

$$T_- \leq 2^j \leq T_+.$$

For the selected-estimator construction, set the external budget to $T_{\max} = T_+$. The candidate pool consisting of the background and powers of two through T_+ therefore contains a good iterate; its size is $s = 2 + \lfloor \log_2 T_+ \rfloor$. In the degenerate branch the background is already a good candidate.

Apply the oracle inequality in Theorem 3 conditionally on training at level $\delta/2$. A sufficient explicit validation condition, by (69), is $n_v \geq 1024\omega^4\epsilon^{-2}\log(8s/\delta)$. Combining its two concentration events with the training event proves the joint $1-\delta$ assertion stated after Theorem 3. More explicitly, substitute the right side of (5) for the candidate minimum in (6), and replace its logarithm by $\log(8s/\delta)$. Thus the guarantee for the selected estimator includes both existence and selection of a good candidate, rather than assuming its existence. This proof uses the multiplicative width of the window.

C.8 Proof of Theorem 3

Early stopping is also studied as regularization in kernel regression (Raskutti et al., 2014). Here we analyze independent validation over a finite candidate set. Conditional on training, all candidates are fixed. Set $\boldsymbol{h}_j = \boldsymbol{b}_j - \boldsymbol{b}^\star$, $r_j = \boldsymbol{h}_j^\top \Sigma \boldsymbol{h}_j$, and $\widehat{r}_j = \|X_v \boldsymbol{h}_j\|_2^2 / n_v$. For $r_j > 0$, the standardized contrasts are sub-Gaussian with scale at most $C_0\omega$. Bernstein's inequality gives

$$\mathbb{P}\{|\widehat{r}_j - r_j| > \epsilon r_j \mid \text{training}\} \le 2\exp(-c n_v \epsilon^2/\omega^4).$$

If $r_j = 0$, the contrast vanishes almost surely. A union bound and the sample-size requirement therefore imply, with failure probability at most $\delta/2$,

$$(1-\epsilon) r_j \le \widehat{r}_j \le (1+\epsilon) r_j \quad (j \le s). \tag{75}$$

Conditional also on validation predictors, the noise condition gives, simultaneously for every j, with failure probability at most $\delta/2$,

$$\begin{aligned}\left|\frac{2}{n_v}\boldsymbol{\varepsilon}_v^\top X_v \boldsymbol{h}_j\right| &\le 2\sigma\sqrt{\frac{2\widehat{r}_j \log(4s/\delta)}{n_v}} \\ &\le \epsilon \widehat{r}_j + \frac{2\sigma^2 \log(4s/\delta)}{\epsilon n_v}.\end{aligned} \tag{76}$$

The second step is $2ab \le \epsilon a^2 + b^2/\epsilon$. Denote the final additive term by a_v, only within this proof. Subtract the common noise square from the validation loss:

$$\widehat{L}_j = \frac{\|\boldsymbol{y}_v - X_v \boldsymbol{b}_j\|_2^2 - \|\boldsymbol{\varepsilon}_v\|_2^2}{n_v} = \widehat{r}_j - \frac{2}{n_v}\boldsymbol{\varepsilon}_v^\top X_v \boldsymbol{h}_j.$$

If $\widehat{j}$ minimizes validation loss, then for every j,

$$\begin{aligned}(1-\epsilon)^2 r_{\widehat{j}} \le (1-\epsilon)\widehat{r}_{\widehat{j}} \le \widehat{L}_{\widehat{j}} + a_v \le \widehat{L}_j + a_v \\ \le (1+\epsilon)\widehat{r}_j + 2a_v \le (1+\epsilon)^2 r_j + 2a_v.\end{aligned} \tag{77}$$

Divide by $(1-\epsilon)^2$ and minimize over j. This proves the general oracle inequality with additive constant 4, provided the absolute constant in its sample-size requirement is sufficiently large. The oracle inequality alone does not assert existence of a good candidate; Theorem 2 and Appendix C.7 supply that assertion for the conservative run. Decreasing ϵ increases the sample requirement and the additive cost.

C.9 Proof of Proposition 4

For fixed $\alpha > 0$, define the potential

$$\Phi_\alpha(\boldsymbol{w}) = \tfrac{1}{4}\sum_e \left[w_e \operatorname{arsinh}\!\left(\frac{w_e}{2\alpha^2}\right) - \sqrt{w_e^2 + 4\alpha^4} + 2\alpha^2 \right]. \tag{78}$$

This potential is a scaled hypentropy (Ghai et al., 2020). The flow endpoint is the unique minimizer of Φ_α over $\widetilde{X}_k \boldsymbol{w} = \widetilde{\boldsymbol{y}}_k$, $P_k^\perp \boldsymbol{w} = \boldsymbol{0}$. Under full background rank, this is equivalent to minimizing $\Phi_\alpha(D_k \boldsymbol{b})$ over $X\boldsymbol{b} = \boldsymbol{y}$, as proved below.

Only within this proof, write $\Phi_\alpha(\boldsymbol{w}) = \sum_e \psi_\alpha(w_e)$, where

$$\psi_\alpha(s) = \tfrac{1}{4}\left[s \operatorname{arsinh}\!\left(\frac{s}{2\alpha^2}\right) - \sqrt{s^2 + 4\alpha^4} + 2\alpha^2 \right].$$

Step 1: the induced geometry. The balanced factor flow and product rule give

$$\dot{\boldsymbol{u}} = -2\boldsymbol{u} \odot \nabla \mathcal{L}_k(\boldsymbol{w}), \quad \dot{\boldsymbol{v}} = 2\boldsymbol{v} \odot \nabla \mathcal{L}_k(\boldsymbol{w}), \quad \frac{d}{dt}(u_e v_e) = 0, \quad u_e v_e = \alpha^2.$$

Consequently $(u_e^2 + v_e^2)^2 = w_e^2 + 4\alpha^4$, and

$$\dot{w}_e = -4\sqrt{w_e^2 + 4\alpha^4}\, \partial_e \mathcal{L}_k(\boldsymbol{w}), \qquad \frac{d}{dt}\nabla\Phi_\alpha(\boldsymbol{w}) = -\nabla \mathcal{L}_k(\boldsymbol{w}), \tag{79}$$

because $\psi'_\alpha(s) = \operatorname{arsinh}(s/(2\alpha^2))/4$ and $\psi''_\alpha(s) = 1/(4\sqrt{s^2 + 4\alpha^4}) > 0$.

Step 2: boundedness and loss convergence. For any feasible comparator $\bar{\boldsymbol{w}}$, define

$$B_{\Phi_\alpha}(\bar{\boldsymbol{w}}, \boldsymbol{w}) = \Phi_\alpha(\bar{\boldsymbol{w}}) - \Phi_\alpha(\boldsymbol{w}) - \langle \nabla\Phi_\alpha(\boldsymbol{w}), \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle.$$

Differentiating in the second argument and applying (79) gives

$$\begin{aligned}\frac{d}{dt} B_{\Phi_\alpha}(\bar{\boldsymbol{w}}, \boldsymbol{w}(t)) &= \langle \nabla\mathcal{L}_k(\boldsymbol{w}), \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle \\ &= -\|\widetilde{X}_k\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k\|_2^2 - \gamma\|P_k^\perp \boldsymbol{w}\|_2^2 = -2\mathcal{L}_k(\boldsymbol{w}). \end{aligned} \tag{80}$$

Also, $\frac{d}{dt}\mathcal{L}_k(\boldsymbol{w}) = -4\sum_e \sqrt{w_e^2 + 4\alpha^4}(\partial_e\mathcal{L}_k(\boldsymbol{w}))^2 \le 0$. Integrating (80) from zero, where $\boldsymbol{w}(0) = \mathbf{0}$, yields

$$2t\mathcal{L}_k(\boldsymbol{w}(t)) \le 2\int_0^t \mathcal{L}_k(\boldsymbol{w}(s))\, ds \le B_{\Phi_\alpha}(\bar{\boldsymbol{w}}, \mathbf{0}) = \Phi_\alpha(\bar{\boldsymbol{w}}).$$

The divergence has the explicit expression

$$B_{\Phi_\alpha}(\bar{\boldsymbol{w}}, \boldsymbol{w}) = \Phi_\alpha(\bar{\boldsymbol{w}}) + \tfrac{1}{4}\sum_e \left[\sqrt{w_e^2 + 4\alpha^4} - 2\alpha^2 - \bar{w}_e \operatorname{arsinh}(w_e/(2\alpha^2))\right]. \tag{81}$$

For fixed $\alpha, \bar{\boldsymbol{w}}$, each summand grows linearly minus at most logarithmically as $|w_e| \to \infty$. Its sublevel sets are bounded. Thus the smooth flow stays bounded, exists globally, and has loss tending to zero. The positive factors recovered from w_e and $u_e v_e = \alpha^2$ remain finite as well.

Step 3: identifying the endpoint. Since $\nabla\Phi_\alpha(\mathbf{0}) = \mathbf{0}$, integrating (79) shows

$$\nabla\Phi_\alpha(\boldsymbol{w}(t)) = -\widetilde{X}_k^\top \int_0^t (\widetilde{X}_k\boldsymbol{w}(s) - \widetilde{\boldsymbol{y}}_k)\, ds - \gamma P_k^\perp \int_0^t \boldsymbol{w}(s)\, ds \in \operatorname{range}(\widetilde{X}_k^\top) + \operatorname{range}(P_k^\perp).$$

This subspace is closed. Every accumulation point is feasible by loss convergence and satisfies the KKT stationarity condition for minimizing Φ_α under the affine constraints. Strict convexity makes this minimizer unique, so the bounded trajectory has exactly one accumulation point and converges to it.

For a consistent $\boldsymbol{w}$, $D_k\boldsymbol{b}(\boldsymbol{w}) = \boldsymbol{w}$. Moreover, (20) shows

$$\widetilde{X}_k\boldsymbol{w} = \widetilde{\boldsymbol{y}}_k \implies X\boldsymbol{b}(\boldsymbol{w}) - \boldsymbol{y} = \sqrt{n}(\widetilde{X}_k\boldsymbol{w} - \widetilde{\boldsymbol{y}}_k) = \mathbf{0}.$$

Conversely, if $X\boldsymbol{b} = \boldsymbol{y}$, then $\boldsymbol{w} = D_k\boldsymbol{b}$ is feasible. The difference between $\boldsymbol{b}$ and its reconstruction lies in $\ker D_k = \operatorname{range}(Q_k)$ and is annihilated by X. Full rank of XQ_k makes that difference zero. This proves the one-to-one correspondence and the coefficient characterization.

Step 4: vanishing initialization. Put $a_\alpha = \log(1/\alpha)$ for $0 < \alpha < 1$. The identity

$$\psi_\alpha(s) = \tfrac{1}{2}a_\alpha|s| + \tfrac{1}{4}\left[|s|\log\frac{|s| + \sqrt{s^2 + 4\alpha^4}}{2} - \sqrt{s^2 + 4\alpha^4} + 2\alpha^2\right]$$

implies

$$\sup_{|s|\le M}\left|\frac{2\psi_\alpha(s)}{a_\alpha} - |s|\right| \longrightarrow 0 \quad \text{for every finite } M. \tag{82}$$

Indeed, $\sqrt{s^2 + 4\alpha^4} \le |s| + 2\alpha^2$ and $|s| \le (|s| + \sqrt{s^2 + 4\alpha^4})/2 \le |s| + \alpha^2$. The expression in brackets is uniformly bounded on bounded intervals, using boundedness of $|s|\log|s|$ near zero.

To bound the minimizers uniformly before invoking this limit, compute the conjugate:

$$\psi_\alpha^*(t) = \frac{\alpha^2}{2}\{\cosh(4t) - 1\}.$$

Fenchel's inequality with $t = \operatorname{sign}(s)a_\alpha/4$ gives $\psi_\alpha(s) \ge a_\alpha|s|/4 - \alpha/4$. For a fixed feasible $\bar{\boldsymbol{w}}$, optimality of $\boldsymbol{w}_\alpha$ implies

$$\frac{a_\alpha}{4}\|\boldsymbol{w}_\alpha\|_1 - \frac{m_k\alpha}{4} \le \Phi_\alpha(\boldsymbol{w}_\alpha) \le \Phi_\alpha(\bar{\boldsymbol{w}}) = \frac{a_\alpha}{2}\|\bar{\boldsymbol{w}}\|_1 + O(1).$$

Hence $\boldsymbol{w}_\alpha$, and its affine reconstruction $\boldsymbol{b}_\alpha$, are bounded. Along any convergent subsequence $\boldsymbol{w}_{\alpha_j} \to \boldsymbol{w}_0$, feasibility is preserved and (82) gives, for every feasible $\bar{\boldsymbol{w}}$,

$$\|\boldsymbol{w}_0\|_1 = \lim_j \frac{2\Phi_{\alpha_j}(\boldsymbol{w}_{\alpha_j})}{a_{\alpha_j}} \le \lim_j \frac{2\Phi_{\alpha_j}(\bar{\boldsymbol{w}})}{a_{\alpha_j}} = \|\bar{\boldsymbol{w}}\|_1.$$

The coefficient correspondence proves the claimed minimum-variation limit. Uniqueness gives convergence rather than only a statement about accumulation points.

A noiseless uniqueness condition. For $S = \operatorname{supp}(D_k \boldsymbol{b}^\star)$, suppose every nonzero $\boldsymbol{h} \in \ker X$ satisfies $\|(D_k \boldsymbol{h})_S\|_1 < \|(D_k \boldsymbol{h})_{S^c}\|_1$. Then

$$\|D_k(\boldsymbol{b}^\star + \boldsymbol{h})\|_1 \geq \|D_k \boldsymbol{b}^\star\|_1 - \|(D_k \boldsymbol{h})_S\|_1 + \|(D_k \boldsymbol{h})_{S^c}\|_1 > \|D_k \boldsymbol{b}^\star\|_1.$$

Thus $\boldsymbol{b}^\star$ is the unique minimum-variation interpolant. It also implies $\ker X \cap \ker D_k = \{\mathbf{0}\}$, since a nonzero vector in that intersection would violate the strict inequality.