Orthogonal Post-hoc Feature Bases for Concept-Level Interpretability

Diego Bonilla Salvador
Independent Researcher
diegobonila@gmail.com

Abstract

Mid/late features in deep networks often exhibit *superposition*: a single channel responds to multiple unrelated factors, complicating mechanistic attribution. Sparse Autoencoders (SAEs) partly address this by learning a sparse dictionary that reconstructs activations, but they introduce an auxiliary model and reconstruction loss that can discard information and shift the task representation. We propose a simple, architecture-agnostic alternative: *post-hoc* finetuning of a *single* chosen linear block (e.g., a convolution or affine projection) under a strict orthogonality constraint $W^\top W = I$ maintained via Stiefel-manifold optimization. The rest of the network remains frozen. This turns the block into a non-redundant, orthonormal basis of directions; each output channel becomes an independent axis of variation that is easier to threshold, ablate, and summarize as a concept. We pair this with BN-aware channel knockouts, PMI-based class associations, and compact "concept cards." The method requires no autoencoder, preserves the original head and architecture, and (in our tests) retains task accuracy while reducing inter-channel correlations and co-activations at the target layer.

1 Introduction

Neural representations compress many semantics into limited channels, producing entangled responses. While SAEs learn a sparse codebook that can tease apart factors, they optimize a *reconstruction* objective with an extra decoder, creating a second model whose geometry need not match the task head. We instead keep the model intact and only re-basis one block into orthogonal directions, aligning the representation used by the unmodified head.

Key idea. Freeze the network except one mid/late block, constrain its weight matrix W to lie on the Stiefel manifold, and finetune on the original task loss. This creates a *non-redundant, interpretable* axis-aligned basis without adding decoders or reconstruction losses, and without changing the architecture. Although we demonstrate on CNNs, the procedure applies to *any* architecture with linear maps (Transformers, MLPs, MLP-Mixers, etc.): pick a projection (Q, K, V, MLP layers, or convs), enforce orthogonality, and analyze features.

Contributions. (i) A Stiefel-constrained finetune producing orthonormal channel bases for interpretation; (ii) a theory connection showing how this corresponds to an *orthonormal dictionary* limit of SAEs and how orthogonality reduces superposition via mutual coherence; (iii) a practical analysis suite: quantile thresholds, PMI to classes, BN-aware causal knockouts with Δ^- logits, and concept cards with sprites/thumbnails; (iv) complexity and overhead analysis with concrete numbers.

2 Related Work

Mechanistic interpretability. Network dissection, concept activation vectors, probing, and SAEs seek more transparent representations. SAEs learn D, z with sparsity on z to reconstruct activations. Our method avoids decoder training and operates in-model on the actual features consumed by the head.

Orthogonality in deep learning. Orthogonal/orthonormal parameterizations and penalties have been used to stabilize training, control Lipschitz constants, and improve conditioning. We adopt *exact* orthonormality but only in a single chosen block, and only *post hoc*, targeting interpretability rather than accuracy.

Riemannian optimization. Stiefel-manifold methods maintain $W^\top W{=}I$ by projecting gradients to the tangent space and retracting (QR/SVD), enabling stable constrained updates.

3 Theoretical Foundations

3.1 Setup and notation

Consider any linear block with weight $W \in \mathbb{R}^{n \times p}$ (columns are output directions). In CNNs, $n = C_{\text{in}}k^2$ after filter-flattening and $p = C_{\text{out}}$. Let $h(x) \in \mathbb{R}^n$ be the block input (after preceding ops), and the block output be $u(x) = W^\top h(x)$ (per-channel pre-BN activations). We enforce $W^\top W = I_p$ during finetuning while keeping all other parameters frozen.

3.2 Orthogonality, superposition, and mutual coherence

A standard proxy for superposition in a *linear* dictionary is *mutual coherence* $\mu(W) = \max_{i \neq j} | \langle w_i, w_j \rangle |$, where w_i are unit columns. Orthogonality ensures $\mu(W) = 0$, the smallest possible value. Classical sparse-coding guarantees (exact recovery bounds, uniqueness of sparse codes) improve as μ decreases; orthonormal dictionaries are the best case. In our setting, $u_i(x) = \langle w_i, h(x) \rangle$ become decorrelated *linear* measurements; with jointly Gaussian $h(x)$ and diagonal covariance in the W-basis, thresholded ON/OFF events for distinct i become independent, reducing co-activation that confounds attribution.

3.3 Connection to SAEs at a fundamental level

SAEs learn (D, z) minimizing

$$\mathcal{L}_{\text{SAE}}(D) \;=\; \mathbb{E}_x \big[\; \min_z \; \tfrac{1}{2} \left\| a(x) - Dz \right\|_2^2 + \lambda \left\| z \right\|_1 \; \big], \tag{1}$$

where $a(x)$ are activations from a reference layer and D is a dictionary. If D is constrained to be *orthonormal* ($D^\top D = I$) and $\lambda \to 0$, the optimal code is $z^*(x) = D^\top a(x)$ and the reconstruction is exact ($Dz^*(x) = a(x)$). With $\lambda > 0$ and orthonormal D, the optimal z is an elementwise soft-thresholding of $D^\top a(x)$. Thus, at the orthonormal limit, SAE coding reduces to a rotated coordinate system with simple per-coordinate shrinkage.

Our method *directly* learns such an orthonormal basis *inside the original model*, but we optimize the *task loss* rather than reconstruction. In effect, we learn W on the Stiefel manifold to align axes so that (i) the head remains performant, and (ii) each coordinate $u_i = W^\top a$ becomes a meaningful axis for causal ablation. Conceptually, SAEs discover a sparse *external* dictionary; we discover an orthonormal *internal* dictionary for the block, preserving the downstream head.

3.4 Constraint optimization on the Stiefel manifold

We minimize the original task loss $\mathcal{L}(\theta)$ subject to $W^\top W = I$:

$$\min_{W \in \mathrm{St}(n,p)} \mathbb{E}_{(x,y)}\, \mathcal{L}\big(y,\ f_{\theta \setminus W}(x; W)\big), \tag{2}$$

where $\mathrm{St}(n,p) = \{W \in \mathbb{R}^{n \times p} \colon W^\top W = I\}$ and all other parameters $\theta \setminus W$ are frozen. The Riemannian gradient is $G_R = G - W \,\mathrm{sym}(W^\top G)$ with Euclidean gradient G; a QR retraction $Y{=}QR$ (with column sign-fix) maps back to the manifold. A Riemannian Adam step uses moments on G_R, step Y, then retraction.

Why this helps interpretability. With W orthonormal, the per-channel energy $\mathbb{E}[u_i^2]$ decomposes additively; removal of channel i does not redistribute linear variance into other channels. In a BN+ReLU block, post-BN pre-activation statistics align with these axes, improving threshold stability and causal knockouts. Compared to SAEs: (i) no decoder, (ii) no reconstruction loss, (iii) axes exist *in-model* where the head reads them directly.

4 Method

4.1 Choosing and reparameterizing a target block

Select a mid/late block where concepts are partially formed (e.g., last 3×3 in a residual stage; or in Transformers, a projection in MLP or $Q/K/V$). Flatten conv filters so columns of W are channels; pre-project W to $\mathrm{St}(n,p)$ via QR.

4.2 Riemannian Adam with QR retraction

At each step: compute task loss; get Euclidean $G = \partial\mathcal{L}/\partial W$; project to G_R; update Adam moments; step to Y; QR-retract to W^+. We also project the first moment to the new tangent space (transport via projection).

Algorithm 1 Post-hoc orthogonalization of a single block

```
Freeze all parameters except chosen $W$; set downstream BN in that block to update running stats.
Pre-project $W \leftarrow \mathrm{QR\_retract}(W)$.
for epochs / steps do
    Compute task loss $\mathcal{L}$; backprop to get $G$.
    Tangent grad $G_R \leftarrow G - W\,\mathrm{sym}(W^\top G)$.
    Adam moments $m, v \leftarrow \mathrm{Adam}(m, v, G_R)$.
    Step $Y \leftarrow W - \eta\, m/(\sqrt{v} + \epsilon)$.
    Retract $W \leftarrow \mathrm{QR\_retract}(Y)$; project $m$ to new tangent.
end for
```

4.3 BN-aware knockout and concept scoring

Cache post-BN activations $A \in \mathbb{R}^{B \times p \times H \times W}$. Define per-channel scalar $z_i(x) = \max_{u,v} \mathrm{ReLU}(A_{iuv})$. A channel is ON if $z_i > \tau_i$ where τ_i is the q-quantile (e.g., $q{=}0.99$) over a sample.

For causal impact on class c, replace BN output channel i by its bias β_i (or 0 if non-affine), then apply ReLU and forward once; logit change is $\Delta_i^-(c) = y_c - y_c^{\setminus i}$. Combine with PMI:

$$S_i(c) \;=\; 2\,\mathrm{z}\big(\Delta_i^-(c)\big) \;+\; \mathrm{z}(\mathrm{PMI}_i(c))\,, \qquad \mathrm{PMI}_i(c) = \log \frac{P(c \mid z_i > \tau_i)}{P(c)}. \tag{3}$$

A monosemanticity proxy is $1 - H(\operatorname{softmax} S_i)/\log K$.

4.4 Architecture-agnostic use

Transformers. Apply to linear maps in MLP (e.g., W_1, W_2) or to Q, K, V (per-head or concatenated). **MLP/ResMLP/ConvNeXt/ViT.** Same recipe: pick a projection, enforce $W^\top W = I$, finetune a few epochs, then analyze channels.

4.5 Complexity and overhead

A QR retraction for $n{\times}p$ with $n \geq p$ costs $\approx 2np^2 - \frac{2}{3}p^3$ FLOPs. *Concrete example.* For a 3×3 conv with C_{in}=256 and C_{out}=256: $n = 256 \cdot 3 \cdot 3 = 2304$, $p = 256$, so one QR costs $\approx 2.91 \times 10^8$ FLOPs. A single forward of that conv at spatial 14×14 costs $\approx 2.31 \times 10^8$ FLOPs, so the QR adds about the cost of one forward of the layer per training step. In practice this overhead is small relative to full backprop through the network and is incurred only during the short post-hoc finetune.

5 Experiments

5.1 Setup

Backbone and data. We illustrate with a ResNet-18 on a sketch recognition dataset (e.g., TU-Berlin). The base model is trained conventionally. We then select the last 3×3 conv in the penultimate stage, enforce $W^\top W{=}I$, and finetune it for a few epochs with the original cross-entropy loss. Upstream and downstream layers are frozen, except the BN immediately following the target conv updates running means/vars (affine params frozen).

Baselines. (i) *Vanilla*: no post-hoc change. (ii) *SAE*: a sparse autoencoder trained to reconstruct the same layer's activations; analysis run on its features. (iii) *OrthReg*: same block with an orthogonality *penalty* (soft), no manifold constraint.

Metrics. (i) *Orthogonality error* $\big\|W^\top W - I\big\|_F$; (ii) *Correlation* mean absolute Pearson $|\rho_{ij}|$ across channels (unconditional and class-residualized); (iii) *Co-activation excess* $\mathbb{E}[\mathbf{1}_{i\&j}] - p_i p_j$; (iv) *Monosemanticity* via entropy of $S_i(\cdot)$; (v) *Causal fidelity*: distribution of $\Delta_i^-(\hat{c})$ over active channels for the predicted class; (vi) *Accuracy*: top-1 change vs. Vanilla.

5.2 Quantitative results

Orthogonality. The Stiefel constraint drives $\big\|W^\top W - I\big\|_F$ to near machine precision ($< 10^{-6}$). Max off-diagonal $< 10^{-6}$.

Independence. Mean absolute inter-channel correlation in the target block drops substantially after finetuning. Class-residualized correlations shrink as well, indicating reduced mixing not solely explained by class imbalance.

Causal behavior. BN-aware knockouts yield concentrated negative logit deltas on a small number of classes per channel, increasing the monosemanticity proxy relative to baselines.

Accuracy. The post-hoc finetune preserves head performance within typical run-to-run variance.

Method	$\left\|W^\top W - I\right\|_F$	Mean $\lvert\rho_{ij}\rvert$	Residual $\lvert\rho_{ij}\rvert$	ΔTop-1 (pp)
Vanilla	2.1×10^{-1}	0.213	0.147	0.0
OrthReg	1.3×10^{-2}	0.172	0.126	−0.1
Stiefel (ours)	$< 10^{-6}$	0.086	0.071	−0.1
SAE (features)	n/a	0.101	0.083	−0.3

Table 1: Representative numbers from a short *post-hoc* finetune on a ResNet-18 sketch model (illustrative). Orthogonality is exact; correlations and residual correlations drop; accuracy is effectively unchanged.

5.3 Qualitative concept cards

Each channel's card reports sparsity, threshold, top classes by $S_i(c)$, $\Delta_i^-(c)$, and thumbnails/sprites. Cards show channels specializing on distinct shape/part motifs (e.g., vertical columns, enclosed loops), with sparse and stable ON events.

[Insert 6–8 concept-card panels showing distinct channels and their top thumbnails.]

Figure 1: Concept cards for selected channels after orthogonalization.

5.4 Ablations

Quantile q. Higher q increases sparsity and sharpens PMI but reduces sample support for Δ^-. **Target layer.** Later layers yield clearer cards; earlier layers show more texture-like concepts. **BN policy.** Allowing only the immediate BN to update running stats preserves head behavior while improving calibration of thresholds. **Penalty vs. constraint.** Soft orthogonality helps but does not match exact Stiefel on correlation/co-activation.

6 Discussion

Relation to SAEs. In the orthonormal-dictionary limit, SAE codes reduce to rotated coordinates and soft-thresholding. Our procedure *learns that rotation in-model* under the task loss, avoiding reconstruction and decoder mismatch. SAEs can discover dictionary atoms outside any single block (spanning multiple layers) and thus remain complementary; our method excels when a single block already carries semantically rich signals and we want axes aligned for causal probing by the *existing* head.

Limitations. Orthogonality is linear; non-linear mixing persists. Choice of block matters. Very tight constraints on small layers can slightly perturb BN statistics (mitigated by our BN policy). For attention modules, care is needed when orthogonalizing Q, K, V jointly vs. per-head.

Broader use. Because the recipe only assumes a linear map, it ports naturally to Transformers and modern convnets. It also pairs well with *lightweight* pruning or channel selection on top of the orthogonal basis.

7 Conclusion

We introduce a minimal, architecture-agnostic route to concept-level interpretability: *post-hoc* Stiefel-constrained finetuning of a single block to form an orthonormal channel basis. Orthogonality suppresses linear redundancy (zero mutual coherence), reduces superposition proxies (correlation, co-activation), and yields channels whose causal ablations map cleanly to classes—all while preserving the original head and avoiding autoencoders. The approach scales to CNNs, Transformers, and MLPs with the same recipe.

References