THE LEARNING GEOMETRY OF DETERMINISTIC VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adding deterministic views of an observation can change a model's predictions without adding information. This makes a standard view ablation difficult to interpret: the input expansion may also change the learning rule. We characterize this intervention for full-rank fixed linear views, each embedded by a trainable linear branch and summed before a shared receiver. Their frame operator $S = A^\top A$ preconditions the effective SGD update and determines the norm induced by branch regularization. Two view systems produce the same update for every loss exactly when their frame operators agree. With matched initialization, equal-S systems follow identical trajectories under SGD and scalar-linear optimizer state, including momentum. The classical Parseval transform $AS^{-1/2}$ therefore supplies a neutral control relative to the parent-only model. Nonidentical equal-S frames match exactly in our experiments, whereas equal-trace frames can separate. One multiscale frame changes subject-mean NLL by $+0.0338$ on Sleep-EDF and -0.1215 on ISRUC; Parsevalization removes both differences. Independent schedule selection retains a Sleep-EDF gap of $+0.0999$ (95% CI $[0.0673, 0.1285]$). AdamW and branchwise nonlinearities break the tested equivalences. The frame operator thus specifies what to control when attributing a deterministic view's effect to the representation.

1 INTRODUCTION

A waveform can be supplied alongside a time–frequency transform, an image alongside edge maps, or a sensor trace alongside a multiscale pyramid. Such views often improve a learner even though they are computed from the observation it already receives. The resulting ablation—with or without the extra view—measures a real performance difference. Explaining that difference requires knowing what the extra branch changed. Figure 1 connects this question to an operator-matched control.

Additional evidence is not the only possible explanation. If $D = T(X)$ is deterministic, observing (X, D) is statistically equivalent to observing X:

$$\sigma(X, D) = \sigma(X), \qquad I(Y; D \mid X) = 0. \tag{1}$$

This is Blackwell equivalence (Blackwell, 1953). The learner can nevertheless respond differently to the two presentations. Even when they represent the same functions, their parameter updates can follow different paths. An input ablation can therefore change the optimization problem seen by the learner without changing the evidence available to it.

We study this effect in an interface where it can be isolated exactly. Fixed linear views are passed through trainable linear branches, added together, and then processed by a shared differentiable receiver. The receiver may be nonlinear; the branch-to-fusion interface is linear. Stack the view maps into a full-column-rank matrix A. Their combined map is $B = WA$, so every view system represents the same set of effective maps as the parent-only model. The difference appears in the update:

$$\Delta B = -\eta G S, \qquad S = A^\top A, \qquad G = \nabla_B \ell, \tag{2}$$

for SGD without weight decay. The frame operator S changes how the effective gradient is applied. It also determines the effective norm induced by Euclidean regularization on the branches. The multiscale construction in Figure 1(b) makes this effect explicit: the added averaging branches increase the gain on low-frequency components of the effective gradient.

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.

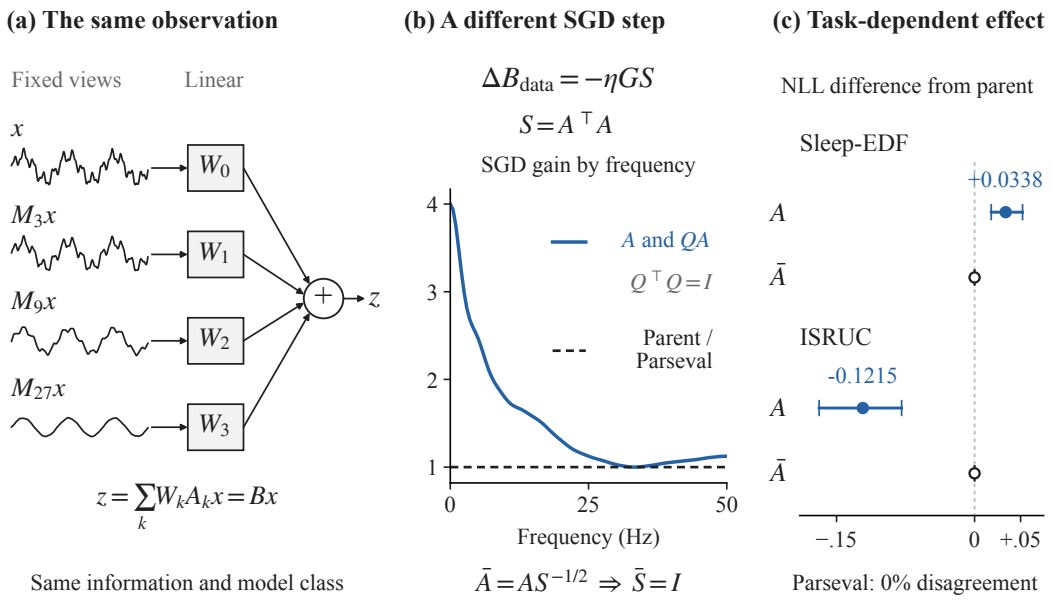

Figure 1: **Deterministic views change the learning rule through their frame operator.** **(a)** The original signal and circular moving averages of widths 3, 9, and 27 pass through trainable linear branches and sum before a shared receiver. The waveforms are schematic; the full-rank construction preserves information and the effective model class. **(b)** Exact frequency-domain gains of the experimental frame $S = A^\top A$, with $G = \nabla_B \ell$. Orthogonal remixing QA changes the views but preserves S and hence the curve. Parsevalization $\bar{A} = AS^{-1/2}$ recovers the parent's unit gain. **(c)** NLL differences from the parent, equally weighted across eight Sleep-EDF and ten ISRUC subjects; whiskers are paired subject-bootstrap 95% confidence intervals. Parsevalization removes both signs of effect: subject-mean gaps are below 6×10^{-9} in magnitude and prediction disagreement is zero. Trajectory equivalence requires the linear additive interface and matched SGD states; the gains in (b) are update multipliers, not convergence rates.

This observation gives two useful comparisons. First, distinct view systems with the same S must follow the same effective trajectory when their initialization, batches, and optimizer state are matched. Such a pair changes the view decomposition while holding its learning geometry fixed. Second, the canonical Parseval transform $\bar{A} = AS^{-1/2}$ sets $\bar{A}^\top \bar{A} = I$, matching the parent-only system. It retains a redundant input presentation while removing its frame-induced change to the update. These comparisons separate the appearance of additional branches from the learning rule they implement.

The ingredients have established origins. Linear adapter–gradient duality (Hennigen et al., 2025) gives the general correspondence between a parameter expansion and a transformed gradient; canonical Parsevalization comes from frame theory (Frank et al., 2002; Janssen & Strohmer, 2002). We use this correspondence to characterize which deterministic input systems constitute the same SGD intervention and to construct the neutral control. The characterization is necessary and sufficient for updates across all losses, and trajectory equivalence extends to scalar-linear optimizer state.

The experiments then ask whether this variable explains the observed differences. Mixing branches while preserving a nonidentity S leaves the entire measured trajectory unchanged. Matching only total energy, or changing eigenspace orientation, does not. On Sleep-EDF and ISRUC, the same multiscale frame has opposite effects on NLL, yet Parsevalization recovers the parent on both. Independent learning-rate and training-horizon selection tests whether scalar retuning removes the difference. Optimizer and branch-nonlinearity tests establish where the control stops being neutral. Together, the analysis and interventions make the deterministic-view comparison precise: within the stated interface, controlling S controls the change in the learning rule.

2 FROM INPUT VIEWS TO LEARNING GEOMETRY

2.1 HOLD INFORMATION AND THE EFFECTIVE MODEL FIXED

Let $x \in \mathbb{R}^d$. Fixed linear views $A_k \in \mathbb{R}^{d_k \times d}$ are stacked as

$$A = \begin{bmatrix} A_1^\top & \cdots & A_K^\top \end{bmatrix}^\top \in \mathbb{R}^{m \times d}, \qquad m = \sum_k d_k,$$

with full column rank. Concatenate trainable branch maps $W_k \in \mathbb{R}^{p \times d_k}$ into $W \in \mathbb{R}^{p \times m}$. Their additive interface is

$$z(x) = \sum_{k=1}^{K} W_k A_k x = WAx = Bx, \qquad B := WA. \tag{3}$$

A shared receiver r_θ acts on z, and the loss depends on the branches only through B: $\mathcal{L}(W, \theta) = \ell(B, \theta)$. This allows nonlinear attention and classification after fusion.

Proposition 1 (Evidence and expressivity equivalence)**.** *If A has full column rank, then observing Ax is statistically equivalent to observing x, and*

$$\{WA : W \in \mathbb{R}^{p \times m}\} = \mathbb{R}^{p \times d}.$$

A left inverse recovers x from Ax, and $W = BA^\dagger$ realizes any effective map B. We can consequently compare different view systems without changing either the observation or the receiver's effective function class. Matching B_0 also makes their initial predictors equal; Appendix B gives the initialization used in the experiments.

2.2 THE EXPANSION CHANGES THE EFFECTIVE UPDATE

Write $G = \nabla_B \ell(B, \theta)$. The chain rule gives $\nabla_W \mathcal{L} = GA^\top$. SGD with learning rate η and coupled ℓ_2 coefficient λ therefore yields

$$W^+ = (1 - \eta\lambda)W - \eta G A^\top, \tag{4}$$

$$B^+ = (1 - \eta\lambda)B - \eta G \underbrace{A^\top A}_{S}. \tag{5}$$

The update closes in the effective coordinates: A enters only through S. For example, stacking two copies of the identity gives $S = 2I$ and doubles the data-gradient term for B at a matched effective state. A nonscalar S instead changes different directions by different amounts. Thus even a lossless expansion can alter the optimization path.

The same matrix specifies the regularizer in effective coordinates. Although many branch parameters realize a given B, their smallest Euclidean norm is fixed by S^{-1}.

Proposition 2 (Induced norm)**.** *For any effective map B,*

$$\min_{W : WA = B} \|W\|_{\mathrm{F}}^2 = \mathrm{tr}(BS^{-1}B^\top), \qquad W_\star = BS^{-1}A^\top. \tag{6}$$

The update metric and induced penalty are therefore linked: changing the frame changes both through the same operator. This explains why a representation comparison with identical scalar hyperparameters can still compare different learning geometries.

2.3 WHICH VIEW SYSTEMS IMPLEMENT THE SAME INTERVENTION?

For full-rank A and $\widetilde{A}$, write $A \sim_{\mathrm{SGD}} \widetilde{A}$ when their one-step effective SGD updates agree for every differentiable loss and matched effective state, at a common positive step size.

Theorem 1 (Frame-operator equivalence)**.** *Consider two full-rank analysis operators A and $\widetilde{A}$ with frame operators S and $\widetilde{S}$. Suppose the training examples, receiver state, scalar hyperparameters, and effective initial states are matched.*

1. *$A \sim_{\mathrm{SGD}} \widetilde{A}$ if and only if $S = \widetilde{S}$.*
2. *If $S = \widetilde{S}$, plain SGD produces identical effective maps and receiver states at every step.*
3. *The trajectory conclusion also holds when branch optimizer state is formed by a shared scalar-linear recurrence from past branch gradients and weights. This class contains momentum, Nesterov momentum, and scalar decoupled weight decay.*

Sufficiency follows because equal effective states give equal gradients in Equation (5). For necessity, equality of the update for every gradient requires $G(S - \widetilde{S}) = 0$ for every G, hence $S = \widetilde{S}$. The universal quantifier matters: a particular trajectory may never probe a direction on which two operators differ. The full proof, including the optimizer-state recurrence, is in Appendix A.

The theorem makes a concrete prediction. Once S is fixed, changing the branch decomposition cannot affect training under these conditions. Changing only its trace, condition number, or eigenvalues need not preserve the update, because the full matrix acts on G.

2.4 Construct a neutral control

The parent-only parameterization has $A = I$. To match its learning geometry, apply the canonical tight-frame construction

$$\bar{A} = AS^{-1/2}, \qquad \bar{A}^\top \bar{A} = I. \tag{7}$$

It is the unique symmetric positive-definite right correction with this property (Frank et al., 2002; Janssen & Strohmer, 2002).

Corollary 1 (Parseval neutrality)**.** *Under the conditions of Theorem 1, initialize the Parsevalized and parent systems with the same effective map and matched effective optimizer state. Their effective maps, receiver states, logits, and predictions are identical at every step.*

This gives a control with an operational interpretation. The expanded system still contains its fixed branches, but its frame-induced update is the same as the parent's. A parent–unnormalized comparison measures the effect of changing S; a parent–Parseval comparison checks that the implementation recovers the identity class.

Two implementation choices can break this comparison. Coordinatewise adaptive optimizer state is not scalar-linear, and a nonlinearity inside each branch prevents the interface from collapsing to WAx. We test these separately below. Numerical error is different: if $\bar{A}^\top \bar{A} = I + E$ with $\|E\|_2 \le \varepsilon$, the SGD update has a perturbation of norm at most $\eta \|G\|_{\mathrm{F}} \varepsilon$. The finite-horizon bound in Theorem 2 propagates this error along the trajectory.

3 Experimental design

The comparisons answer three questions in sequence: whether S isolates the effect of the view decomposition, whether that effect changes task performance, and whether scalar retuning removes it. Within each comparison we match the receiver, effective initialization, batches, and scalar training settings, except for the quantity explicitly varied. All dataset comparisons use three matched seeds.

Frames and operator controls. The parent uses the identity. The multiscale STANDARD frame stacks it with circular moving averages of widths 3, 9, and 27. TRACE divides this frame by $\sqrt{\mathrm{tr}(S)/d}$, matching total energy; PARSEVAL applies $S^{-1/2}$, matching the full operator. Starting from a non-tight frame, two orthogonal mixing constructions preserve the same nonidentity S while changing the branch or coordinate decomposition. Other controls vary the spectrum, rotate eigenspaces at fixed eigenvalues, or match trace while changing the spectrum. Appendices B and D specify the operators.

EEG tasks and receiver. Sleep-EDF (Goldberger et al., 2000; Kemp et al., 2000) uses Fpz–Cz EEG at 100 Hz; ISRUC (Khalighi et al., 2016) uses C3–A2 resampled from 200 to 100 Hz. Each 30 s epoch supplies 29 overlapping windows of width 200 and hop 100. Labels are Wake, N1, N2, N3, and REM. A one-layer transformer receives the fused tokens: width 48, four heads, FFN width 96, mean pooling, and no dropout. The loss is inverse-frequency weighted cross-entropy. The matched EEG study uses SGD with learning rate 0.01, weight decay 10^{-4}, batches of 512, and 10 passes on Sleep-EDF or 20 on ISRUC. The operator and schedule-selection studies use batches of 1024; settings remain matched across arms within each study.

Sleep-EDF has 40 fit subjects and eight evaluation subjects. ISRUC uses five fixed folds, each fitting eight subjects and evaluating the other two. Its recordings use median/MAD scaling, clipping to $[-10, 10]$, and a sleep-period crop. The preprocessing and subject roles are given in Appendix C.

Independent schedule selection. Each Sleep-EDF arm selects its learning rate from $\{0.0025, 0.005, 0.01, 0.02, 0.04\}$ and its pass count from $\{1, 2, 5, 10, 20, 30, 40\}$. Selection uses eight tuning subjects within the fit pool, with the other 32 used for fitting. The selected schedule is refit on all 40 before evaluation on the unchanged eight-subject set. That evaluation set is not used for schedule selection.

Metrics and uncertainty. NLL and Brier score measure probabilistic performance; accuracy and macro-F1 report classification performance. Prediction disagreement and effective-map deviation test trajectory equivalence directly. NLL contrasts and their 95% intervals use paired subject bootstraps with equal subject weight. Aggregate metrics and subject-mean contrasts are reported separately.

Boundary and numerical checks. A 100-step synthetic-input test compares SGD, momentum, Nesterov, scalar decoupled decay, and AdamW (Loshchilov & Hutter, 2019). Further tests vary the Parseval residual from 10^{-4} to 10^{-1} or insert leaky-ReLU with slopes 1, 0.95, and 0.5 before fusion. A CIFAR-100 replication (Krizhevsky, 2009) uses 8-by-8 RGB patches, 16 tokens, a two-layer width-96 transformer, three seeds, and 20 passes. Construction timings cover dense and structured frames. These checks and their complete metrics appear in Appendices F–H.

4 Results

4.1 Equal frame operators give identical trajectories

We first change the view system without changing its frame operator. Orthogonal branch and coordinate mixing preserve the same nonidentity S to relative errors of 2.61×10^{-15} and 2.78×10^{-15}. Across three seeds and every recorded step, both constructions have zero effective-trajectory gap from the base frame and zero prediction disagreement. Their NLL, Brier score, accuracy, and macro-F1 are identical. The changed decompositions therefore have no measured effect once the learning geometry is held fixed.

Controls that preserve less than the full matrix separate. Equal-trace frames with isotropic, ramp, bimodal, and spiked spectra disagree with the same-S reference on 11.5–16.0% of predictions. Rotating a fixed condition-8 spectrum also changes outcomes. Figure 2 places these controls beside the equal-S pairs: matching the operator recovers identity, while matching a scalar summary does not. NLL can move in either direction when S changes, so the equivalence does not imply an ordering by condition number.

4.2 One frame has opposite effects across datasets

The operator tests establish the mechanism. The matched EEG comparison shows its consequence for predictions. The multiscale frame worsens Sleep-EDF subject-mean NLL by 0.0338 (95% CI $[0.0181, 0.0521]$) and improves ISRUC by 0.1215 (CI $[-0.1689, -0.0793]$), changing 14.0% and 24.3% of predictions, respectively (Table 1). Macro-F1 moves in the same directions: from 0.5724 to 0.5621 on Sleep-EDF and from 0.5107 to 0.5633 on ISRUC.

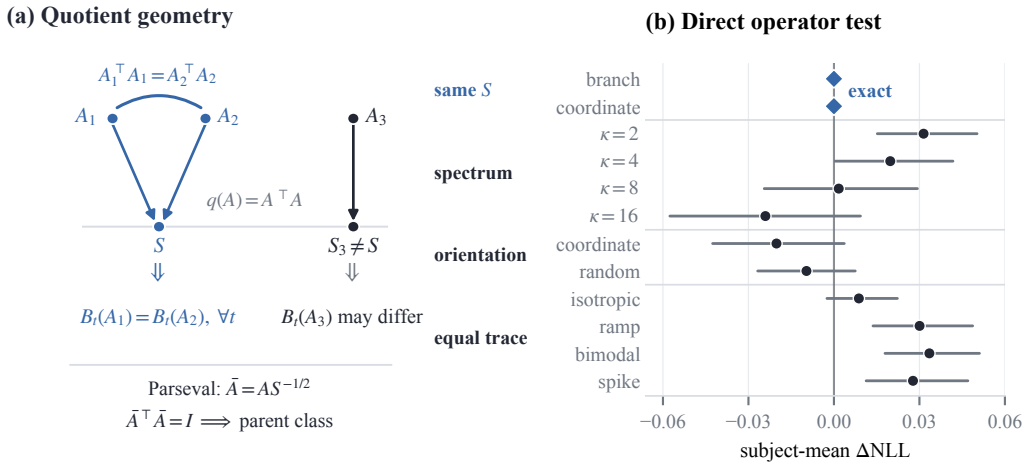

Figure 2: **The frame operator defines the controlled comparison.** **(a)** For the linear interface and matched training states of Theorem 1, equal $S = A^\top A$ gives identical effective trajectories. Equality of S is also necessary for update equivalence across all losses. Parsevalization maps a full-rank frame to $S = I$, the parent class. **(b)** Equal-S pairs match exactly on Sleep-EDF; changing spectrum or orientation, or matching trace alone, can change NLL. Points are equal-subject mean NLL differences from the same-S reference; intervals are paired subject-bootstrap 95% CIs.

Trace normalization retains 75.4% of the original disagreement on Sleep-EDF and 63.8% on ISRUC, with nonzero NLL gaps on both. Parsevalization instead recovers the parent: disagreement is zero, and NLL differs by less than 6×10^{-9} and 2×10^{-9}, respectively. The same control removes both the helpful and the harmful effect. The sign of the performance change depends on the task; its removal follows from matching the update geometry.

Table 1: **The same frame has opposite effects; Parsevalization recovers PARENT on both datasets.** Lower is better for NLL and Brier; higher is better for accuracy and macro-F1. ΔNLL is the equal-subject difference from PARENT, with paired subject-bootstrap 95% CIs; disagreement is also relative to PARENT. Aggregate NLL and subject-mean ΔNLL use different weighting.

Data	Frame	NLL	Brier	Acc.	F1	Disagr.	ΔNLL [95% CI]
Sleep-EDF	PARENT	.8971	.4875	.6293	.5724	—	—
	STANDARD	.9282	.5041	.6193	.5621	13.96%	+.0338 [.0181,.0521]
	TRACE	.9125	.4969	.6219	.5650	10.52%	+.0189 [.0058,.0315]
	PARSEVAL	.8971	.4875	.6293	.5724	0.00%	.0000
ISRUC	PARENT	1.0796	.5703	.5619	.5107	—	—
	STANDARD	.9567	.5144	.6031	.5633	24.34%	−.1215 [−.1689,−.0793]
	TRACE	.9917	.5293	.5985	.5553	15.52%	−.0866 [−.1117,−.0662]
	PARSEVAL	1.0796	.5703	.5619	.5107	0.00%	.0000

4.3 THE DIFFERENCE SURVIVES INDEPENDENT SCHEDULE SELECTION

A different frame may simply require a different learning rate or more training. We test this alternative by selecting each arm's learning rate and pass count independently on the internal tuning subjects. The parent selects learning rate 0.04 and 30 passes; the multiscale frame selects 0.005 and 30 passes.

After refitting, evaluation NLL is 0.7780 for the parent and 0.8799 for the multiscale frame. The equal-subject difference is 0.0999 (95% CI $[0.0673, 0.1285]$), with a positive gap for all eight subjects (Appendix E). Brier, accuracy, and macro-F1 also favor the parent: $0.4248/0.6819/0.6221$ versus $0.4725/0.6446/0.5928$. Thus the difference persists after arm-specific selection over the

evaluated learning rates and training horizons. It cannot be attributed solely to imposing the original common scalar schedule.

4.4 Optimizer state and branch nonlinearities break the control

The exact result uses both a linear interface and scalar-linear optimizer state. We change these conditions separately to identify their roles.

Optimizer state. On synthetic inputs and labels, SGD, momentum, Nesterov, and scalar decoupled decay keep both the Parseval–parent and same-S logit gaps within 1.93×10^{-4} over 100 matched steps (Table 2). AdamW produces gaps of 0.244 and 0.286. Its coordinatewise second-moment update acts before the branches are collapsed into B, so the scalar-linear recurrence used in the theorem no longer applies. In this test, matching S alone is insufficient.

Table 2: **Optimizer state determines whether the control remains neutral.** Maximum absolute logit gap over 100 matched steps on synthetic inputs and labels. AdamW breaks both equal-S comparisons in this test.

Optimizer	Parseval–Parent	Same-S pair
SGD	1.42×10^{-4}	1.36×10^{-4}
Momentum	1.49×10^{-4}	1.31×10^{-4}
Nesterov	1.57×10^{-4}	1.49×10^{-4}
Scalar decoupled decay	1.92×10^{-4}	1.65×10^{-4}
AdamW	2.44×10^{-1}	2.86×10^{-1}

Branch nonlinearity. With an identity activation before fusion, the Parseval control reproduces the linear result. A leaky-ReLU slope of 0.95 produces 0.40% prediction disagreement; a slope of 0.5 produces 7.16%, with NLLs of 0.9258 for the parent and 0.9111 for the Parsevalized branches. The latter comparison changes the function interface: independently activated branches no longer reduce to WAx. Across these tested slopes, disagreement grows as the activation moves away from linearity.

4.5 Numerical precision and an image-domain replication

Approximate Parsevalization produces a smaller, continuous departure. At 200 SGD steps, residuals $10^{-4}, 10^{-3}, 10^{-2}, 0.05, 0.1$ give effective-map gaps of 1.83×10^{-6}, 1.58×10^{-5}, 1.57×10^{-4}, 7.87×10^{-4}, and 1.57×10^{-3}. The observed near-linear dependence is consistent with Theorem 2; the complete numerical results are in Appendix F.

On CIFAR-100, the base frame and a nonidentical equal-S construction again have zero effective-trajectory gap, zero test-prediction disagreement, and identical metrics across three seeds. Parseval and parent stay within 5.24×10^{-6} in effective-map norm and 2.9×10^{-6} in aggregate NLL. This is an image-domain test of the same dynamical prediction with a patch transformer; Appendix G reports its classification metrics. Appendix H gives construction timings and residuals for dense, Fourier-diagonal, blockwise, and iterative Parsevalization.

5 What a deterministic-view ablation measures

In the analyzed interface, a comparison between the parent and an unnormalized view system changes the effective update metric and induced penalty together. A gain is compatible with a useful preconditioner; a loss is compatible with an unsuitable one. The opposite EEG outcomes illustrate both possibilities. Neither requires a change in the information or effective function class.

The frame operator turns this interpretation into a usable control. An equal-S pair holds the learning geometry fixed while changing the decomposition. The Parseval–parent pair then checks the identity class. Together, they provide a reference for interpreting the unnormalized comparison. If the Parseval pair diverges beyond numerical tolerance, the assumptions or their implementation need to be checked before the difference is attributed to branch semantics.

The relevant checks are specific. Effective initialization and data order must match; coordinatewise adaptive moments require additional state information; and nonlinearities before fusion alter the function class. Rank-deficient analysis can discard information, while a learned frontend changes A during training. These settings require quantities beyond the fixed S studied here. Within the stated setting, the theory determines which trajectories must agree, and the experiments show that changing their common geometry can have either sign of performance effect.

6 Related work

Reparameterization and gradient transformations. Optimization can depend on parameterization even when the represented functions agree (Neyshabur et al., 2015; Gunasekar et al., 2018; Raiko et al., 2012). Hennigen et al. (2025) establish trajectory duality between fixed linear gradient transformations and additive adapters, including stateful optimizers. The connection to our setting is exact:

$$\mathrm{vec}(B) = (A^\top \otimes I_p)\,\mathrm{vec}(W). \tag{8}$$

Their transform is $A \otimes I_p$ in these coordinates. Our analysis uses its SGD specialization to identify equal interventions across input view systems and to build an operator-matched comparison. Linear data augmentation and dropout also induce model-dependent biases (Wu et al., 2020; Helmbold & Long, 2015); here the controlled quantity is the fixed frame at the additive interface.

Frame normalization and frontends. Canonical tight-frame constructions and approximations are classical (Frank et al., 2002; Janssen & Strohmer, 2002). Nonstationary Gabor frames provide structured time–frequency systems (Balazs et al., 2011), while scattering and learnable frontends encode signal structure (Bruna & Mallat, 2013; Zeghidour et al., 2021). Parseval constraints also appear in neural-network Lipschitz control (Cisse et al., 2017). We use canonical Parsevalization as a trajectory-neutral input control: its role follows from the induced update metric rather than a robustness objective or a claim of additional signal.

Information and the learner. Blackwell equivalence concerns the evidence available for decisions (Blackwell, 1953). Usable-information theory instead restricts the predictive model class (Xu et al., 2020). Our comparison fixes both the information and the effective function class, then studies the optimization rule that selects a predictor within it. This distinction is what makes the operator intervention identifiable.

7 Conclusion

Fixed linear views can change learning while preserving both information and the effective model. At an additive linear interface, $S = A^\top A$ specifies that change for SGD and determines which view systems are universally update-equivalent. Canonical Parsevalization supplies the parent-neutral control under matched training conditions. The equal-S comparisons, opposite EEG effects, and independent schedule selection connect this geometry to observed predictions. For deterministic-view ablations in this regime, the frame operator is therefore a concrete part of the experimental design.

Reproducibility statement

The supplement specifies analysis operators, initialization, optimizer recurrences, subject roles, preprocessing, model architecture, selection grids, evaluation counts, and bootstrap units. The accompanying result files contain per-seed and per-subject metrics, trajectory checks, Parseval residuals, and construction timings.

AI Use Statement

In this work, we used generative AI tools to help implement parts of our baseline methods. We did not use generative AI tools to formulate hypotheses, design experiments, develop theoretical results,

write proofs, or interpret results; generating synthetic data, translation, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability and to help format references. All AI-generated code was reviewed, tested, and verified against reference implementations by two of the authors, and all edited text was checked by the authors for accuracy. The authors take full responsibility for the content of this paper, including any text, code, or claims produced with AI assistance.

REFERENCES

Peter Balazs, Monika Dörfler, Florent Jaillet, Nicki Holighaus, and Gino Angelo Velasco. Theory, implementation and applications of nonstationary Gabor frames. *Journal of Computational and Applied Mathematics*, 236(6):1481–1496, 2011. doi: 10.1016/j.cam.2011.09.011.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. doi: 10.1109/TPAMI.2012.230.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863, 2017. URL https://proceedings.mlr.press/v70/cisse17a.html.

Michael Frank, Vern I. Paulsen, and Terry R. Tiballi. Symmetric approximation of frames and bases in Hilbert spaces. *Transactions of the American Mathematical Society*, 354(2):777–793, 2002. doi: 10.1090/S0002-9947-01-02838-0.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

David P. Helmbold and Philip M. Long. On the inductive bias of dropout. *Journal of Machine Learning Research*, 16(105):3403–3454, 2015. URL https://www.jmlr.org/papers/v16/helmbold15a.html.

Lucas Torroba Hennigen, Hunter Lang, Han Guo, and Yoon Kim. On the duality between gradient transformations and adapters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23030–23046. PMLR, 2025. URL https://proceedings.mlr.press/v267/hennigen25a.html.

Augustus J. E. M. Janssen and Thomas Strohmer. Characterization and computation of canonical tight windows for Gabor frames. *Journal of Fourier Analysis and Applications*, 8(1):1–28, 2002. doi: 10.1007/s00041-002-0001-x.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Jos J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. doi: 10.1016/j.cmpb.2015.10.013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Tapani Raiko, Harri Valpola, and Yann LeCun. Deep learning made easier by linear transformations in perceptrons. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 924–932, 2012. URL https://proceedings.mlr.press/v22/raiko12.html.

Sen Wu, Hongyang Zhang, Gregory Valiant, and Christopher Ré. On the generalization effects of linear transformations in data augmentation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10410–10420, 2020. URL https://proceedings.mlr.press/v119/wu20g.html.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jM76BCb6F9m.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Since A has full column rank, $A^\dagger A = I$ and hence $x = A^\dagger(Ax)$. Thus x and Ax generate the same sigma-algebra. For any $B \in \mathbb{R}^{p\times d}$, choosing $W = BA^\dagger$ gives $WA = B$. Conversely every WA is a parent-space linear map. This proves both claims.

A.2 PROOF OF PROPOSITION 2

The constraint $WA = B$ has general solution

$$W = BA^\dagger + N(I - AA^\dagger)$$

for arbitrary N. The two terms are orthogonal in Frobenius inner product, so the minimum is attained at $W_\star = BA^\dagger = BS^{-1}A^\top$. Therefore

$$\|W_\star\|_{\mathrm{F}}^2 = \mathrm{tr}\big(BS^{-1}A^\top AS^{-1}B^\top\big) = \mathrm{tr}(BS^{-1}B^\top).$$

A.3 PROOF OF THEOREM 1

For plain SGD, the chain rule gives $\nabla_W \mathcal{L} = GA^\top$. Multiplication of the branch update by A yields Equation (5). If $S = \widetilde{S}$, the one-step effective updates agree for every G. Conversely, universal equality implies $G(S - \widetilde{S}) = 0$ for every $G \in \mathbb{R}^{p\times d}$. Choosing rows of G as standard basis vectors gives $S = \widetilde{S}$, proving the first claim.

For the second claim, if $S = \widetilde{S}$ and the effective states match, the fused representations, receiver gradients, and G match. The next effective states therefore match, and induction proves equality at every step.

For the third claim, let each branch-state component obey

$$Q_{t+1} = \alpha_t Q_t + \beta_t \nabla_{W_t}\mathcal{L} + \gamma_t W_t$$

with coefficients shared across branch coordinates, allowing a finite stack of such states. Its effective counterpart $R_t = Q_t A$ obeys

$$R_{t+1} = \alpha_t R_t + \beta_t G_t S + \gamma_t B_t.$$

All recurrences close in (B_t, R_t, θ_t) through S alone. Induction again gives equality. Momentum and Nesterov are obtained by the usual scalar recurrences; scalar decoupled weight decay contributes a scalar multiple of B_t.

A.4 PROOF OF COROLLARY 1

Since

$$\bar{A}^\top \bar{A} = S^{-1/2}A^\top AS^{-1/2} = I,$$

the Parsevalized system and the parent analysis operator I have the same frame operator. Theorem 1 applies after matching the effective initialization and optimizer state.

A.5 FINITE-HORIZON STABILITY FOR APPROXIMATE PARSEVALIZATION

For plain SGD, the perturbation enters the effective-map update directly. The following bound makes the numerical comparison used in the main text explicit.

Theorem 2 (Finite-horizon stability under SGD)**.** *Let $U_t = (B_t, \theta_t)$ collect the effective map and receiver parameters, and let their parent SGD update F_t satisfy $\|F_t(U) - F_t(V)\| \le L_t\|U - V\|$. Suppose $\|\nabla_B \ell(U)\|_{\mathrm{F}} \le G_t$ along the compared trajectories. If $\bar{A}^\top \bar{A} = I + E$ with $\|E\|_2 \le \varepsilon$, both systems use the same scalar settings and start at the same effective state, then*

$$\|\widetilde{U}_T - U_T\| \le \varepsilon \sum_{t=0}^{T-1} \eta_t G_t \prod_{j=t+1}^{T-1} L_j. \tag{9}$$

Proof. Let $\Delta_t = \widetilde{U}_t - U_t$. The approximate update can be written as the parent map plus a perturbation supported on the effective-map component:

$$\widetilde{U}_{t+1} = F_t(\widetilde{U}_t) + R_t, \qquad \|R_t\| \le \eta_t \left\|\nabla_B \ell(\widetilde{U}_t)\right\|_{\mathrm{F}} \|E\|_2 .$$

The Lipschitz condition and the assumptions give

$$\|\Delta_{t+1}\| \le L_t \|\Delta_t\| + \eta_t G_t \varepsilon.$$

Because $\Delta_0 = 0$, repeated substitution yields Equation (9). If the output map is L_{out}-Lipschitz in U, the corresponding output deviation is at most L_{out} times the same bound.

B Analysis operators and matched initialization

For EEG, all operators act on one 200-sample token. Circular boundary conditions make the constructions deterministic and remove padding choices. Table 3 lists the main arms.

Table 3: Fixed EEG analysis systems. P_j are disjoint DCT-band projectors, M_w is circular moving-average smoothing of width w, and $S_{\mathrm{std}} = A_{\mathrm{std}}^\top A_{\mathrm{std}}$.

Name	Stacked analysis operator	Frame operator
Parent	I	I
OrthogonalBands	$[P_1^\top, P_2^\top, P_3^\top, P_4^\top]^\top$	$\sum_j P_j = I$
Standard	$[I^\top, M_3^\top, M_9^\top, M_{27}^\top]^\top$	S_{std}
Trace	$A_{\mathrm{std}}/\sqrt{\mathrm{tr}(S_{\mathrm{std}})/d}$	$dS_{\mathrm{std}}/\,\mathrm{tr}(S_{\mathrm{std}})$
Parseval	$A_{\mathrm{std}} S_{\mathrm{std}}^{-1/2}$	I

For every comparison, we initialize in effective coordinates. Given a common B_0, the minimum-norm branch initialization for an operator A is $W_0 = B_0 S^{-1} A^\top$, which satisfies $W_0 A = B_0$. Receiver parameters and data-order seeds are identical across arms. Same-S branch mixing uses an orthogonal map on the stacked analysis coordinates. Same-S coordinate mixing is constructed together with the corresponding branch reparameterization so that the effective initial state and $A^\top A$ are preserved.

C Data, roles, and evaluation

Sleep-EDF. We use the Sleep-EDF Expanded sleep-cassette recordings distributed through PhysioNet (Goldberger et al., 2000; Kemp et al., 2000). Fpz–Cz EEG is sampled at 100 Hz and divided into 30 s epochs. Each epoch contributes 29 overlapping 200-sample tokens with hop 100. Movement and unknown labels are excluded; the remaining stages are mapped to Wake, N1, N2, N3, and REM, with N3 and N4 merged. Forty prespecified subjects form the fit pool:

$$1, 2, 5, 9, 11, 14, 15, 16, 19, 20, 23, 25, 27, 29, 30, 31,$$
$$34, 35, 36, 37, 41, 43, 44, 48, 49, 50, 52, 53, 55, 59, 62, 64,$$
$$65, 66, 70, 71, 72, 74, 80, 82.$$

The evaluation subjects are $\{0, 4, 6, 21, 40, 56, 63, 73\}$. The matched study fits on all 40 subjects. For schedule selection, the first 32 subjects above are used for fitting and the final eight for tuning; the selected schedule is then refit on all 40 before one evaluation on the unchanged eight-subject set. None of the experiments reported here uses the additional reserved Sleep-EDF partitions.

ISRUC. We use ISRUC-Sleep Cohort III (Khalighi et al., 2016). C3–A2 is resampled from 200 to 100 Hz. Each recording is centered by its median, scaled by median absolute deviation, and clipped to $[-10, 10]$. Labels use the same five-stage map. We retain the interval from one hour before the first non-Wake epoch to one hour after the last non-Wake epoch and require eligible 21-epoch context when constructing rows; the receiver itself consumes the center 30 s epoch. Five fixed folds hold out subjects $(1, 2), (3, 4), (5, 6), (7, 8), (9, 10)$ in turn. Each fold fits the remaining eight subjects and evaluates each held-out subject once per seed.

CIFAR-100. The controlled image replication uses all 50,000 official training images and 10,000 official test images (Krizhevsky, 2009). Channel means and standard deviations are computed on training data. Each 32-by-32 RGB image is split into sixteen nonoverlapping 8-by-8 patches, giving coordinate width 192. The image frame stacks identity with circular 3-by-3 and 5-by-5 blur operators; an orthogonal coordinate construction preserves its nonidentity S. Three fixed seeds are trained for 20 passes. The official test set is read once per seed.

D COMPLETE OPERATOR CONTROLS

Table 4 reports aggregate Sleep-EDF metrics for the operator-discrimination study. The base and both same-S constructions are identical in every metric. Condition-number, orientation, and equal-trace arms demonstrate that neither a scalar energy budget nor a single spectral summary substitutes for the matrix S.

Table 4: Operator controls averaged across three seeds. Differences are relative to SAMESBASE.

Arm	NLL	Brier	Acc.	F1	ΔNLL [95% CI]
SAMESBASE	.9457	.5065	.6108	.5629	—
SAMESBRANCHMIX	.9457	.5065	.6108	.5629	.0000 [.0000,.0000]
SAMESCOORDINATEMIX	.9457	.5065	.6108	.5629	.0000 [.0000,.0000]
COND2	.9816	.5237	.5986	.5492	+.0315 [.0152,.0502]
COND4	.9714	.5185	.6031	.5529	+.0198 [.0004,.0418]
COND8	.9553	.5103	.6099	.5586	+.0017 [−.0245,.0293]
COND16	.9308	.4987	.6207	.5679	−.0241 [−.0576,.0093]
ORIENT8COORD.	.9343	.4994	.6200	.5707	−.0202 [−.0426,.0036]
ORIENT8RANDOM	.9421	.5027	.6152	.5671	−.0096 [−.0267,.0075]
TRACEISOTROPIC	.9576	.5125	.6081	.5591	+.0088 [−.0025,.0223]
TRACERAMP	.9805	.5231	.5994	.5499	+.0301 [.0137,.0487]
TRACEBIMODAL	.9830	.5245	.5980	.5485	+.0335 [.0179,.0511]
TRACESPIKE	.9773	.5219	.6013	.5515	+.0277 [.0113,.0470]

E LONG-HORIZON SUBJECT RESULTS

Table 5 gives the per-subject NLL difference STANDARD−PARENTafter independent schedule selection and refitting. All eight gaps are positive.

Table 5: Long-horizon Sleep-EDF per-subject NLL gap.

Subject	0	4	6	21	40	56	63	73
ΔNLL	.1284	.1176	.1583	.0843	.1232	.0413	.0190	.1270

F BOUNDARY DETAILS

F.1 APPROXIMATE PARSEVALIZATION

Table 6: Parseval residual and trajectory deviation after 200 SGD steps.

$\left\|\bar{A}^\top \bar{A} - I\right\|_2$	Effective-map gap	Maximum logit gap
9.31×10^{-9}	8.35×10^{-7}	8.13×10^{-5}
1.00×10^{-4}	1.83×10^{-6}	6.78×10^{-5}
1.00×10^{-3}	1.58×10^{-5}	1.18×10^{-4}
1.00×10^{-2}	1.57×10^{-4}	3.56×10^{-4}
5.00×10^{-2}	7.87×10^{-4}	1.67×10^{-3}
1.00×10^{-1}	1.57×10^{-3}	3.32×10^{-3}

F.2 BRANCHWISE NONLINEARITIES

Table 7: Boundary test with leaky-ReLU before additive fusion. Slope one is the linear interface covered by the theorem.

Slope	Parent NLL	Parseval NLL	Disagreement	Final effective gap range
1.00	.9342	.9342	.0000	numerical zero
.95	.9344	.9353	.0040	.0064–.0084
.50	.9258	.9111	.0716	.1066–.1152

G CIFAR-100 CONTROLLED REPLICATION

This experiment checks the trajectory prediction on image patches with a small transformer. The parent reaches 9.597% accuracy and macro-F1 0.06704; the operator comparison concerns agreement between matched systems at this performance level. Training, test-set use, and the frame construction are specified in Appendix C.

Table 8: CIFAR-100 aggregate test metrics across three seeds.

Arm	NLL	Brier	Accuracy	Macro-F1
IMAGEPARENT	3.994916	.967371	.09597	.06704
IMAGESTANDARD	3.997073	.967203	.09663	.06807
IMAGESAMES	3.997073	.967203	.09663	.06807
IMAGETRACE	3.998733	.967300	.09677	.06817
IMAGEPARSEVAL	3.994913	.967371	.09600	.06708

H CONSTRUCTION ROUTES

Structured operators avoid a generic dense inverse square root. Dense construction reaches dimension 4096 in $0.218\,\mathrm{s}$, a Fourier-diagonal route handles dimension 16384 in $0.150\,\mathrm{s}$, and block-64 construction handles that width in $0.158\,\mathrm{s}$. Ten Newton–Schulz steps give relative residuals between 2.04×10^{-4} and 1.46×10^{-3} through dimension 1024. The table reports one-shot timings and the corresponding residuals.

Table 9: One-shot construction measurements on the experiment host. Residual is relative Frobenius error in the resulting Parseval operator. Fourier entries report finite construction rather than a dense residual.

Route	Dimension	Seconds	Relative residual
Dense eigendecomposition	256	.0055	6.50×10^{-4}
Dense eigendecomposition	1024	.0167	7.93×10^{-4}
Dense eigendecomposition	4096	.2176	1.38×10^{-3}
Newton–Schulz, 10 steps	256	.00038	2.04×10^{-4}
Newton–Schulz, 10 steps	512	.00046	1.28×10^{-3}
Newton–Schulz, 10 steps	1024	.00172	1.46×10^{-3}
Blockwise, block 64	1024	.0101	7.53×10^{-4}
Blockwise, block 64	4096	.0400	7.53×10^{-4}
Blockwise, block 64	16384	.1583	7.53×10^{-4}
Fourier-diagonal	4096	.0261	finite
Fourier-diagonal	16384	.1498	finite