Grade Discovery as an Identifiability Diagnostic for Clifford Valued Features

Anonymous authors
Paper under double-blind review

Abstract

Clifford valued models can represent scalars, vectors, bivectors and pseudoscalars in one algebra, but current models usually assume the geometric type of each feature is known prior to training. This assumption is not always identifiable from the data. For example, rotation only evidence makes scalar and pseudoscalar transformation laws indistinguishable, whereas reflections reveal the difference. We study grade discovery as an identifiability diagnostic, where grade means the Clifford algebra type of a feature channel, such as scalar, vector, bivector or pseudoscalar. The method estimates this type from paired observations before and after known transformations, assigns a soft weight over candidate grades with the same coordinate dimension and fits this weight with a least-squares equivariant prediction loss. We prove that the loss recovers the correct type when the observed transformations separate the candidate transformation laws and that it remains uninformative when those laws agree on all observed transformations. In controlled three-dimensional experiments, rotation only evidence keeps the true type weight at one half, while rotations with reflections assign weight near one to the true type. Reflection frequency and loss landscape experiments show how often parity revealing transformations must appear and why they remove the flat ambiguity. Stress tests further show that inaccurate orthogonal transformation matrices and weak variance in separating directions reduce the diagnostic evidence. Finally, a differentiable soft gate and a minimal trainable prediction module show that the diagnostic can be optimized jointly with another parameter, although this experiment is not evidence of performance in full Clifford neural networks.

1 Introduction

Geometric machine learning often begins by assigning a transformation law to every feature. A temperature is commonly treated as a scalar, a velocity as a vector, an oriented area as a bivector and a signed volume as a pseudoscalar. Once these choices are made, equivariant neural networks can process the resulting features in a way that respects rotations, translations, reflections and other symmetries (Cohen & Welling, 2016; Bronstein et al., 2021; Weiler & Cesa, 2019; Geiger & Smidt, 2022). Despite these advantages, the geometric typing of features is largely treated as an external preprocessing step (the architecture relies on domain experts to map raw coordinates to their corresponding geometric semantics prior to training).

This constraint is especially noticeable in Clifford and Geometric Algebra (GA) models, where Clifford neural layers use multivectors to couple scalar, vector and higher grade fields for physical modelling (Brandstetter et al., 2023). For instance, architectures like Geometric Clifford Algebra Networks and Clifford Group Equivariant Neural Networks explicitly depend on this algebra to parameterize group actions and equivariant maps (Ruhe et al., 2023b;a), while the GA Transformer requires a unified Projective Geometric Algebra (PGA) representation to process diverse geometric object types within a single framework (Brehmer et al., 2023). Building on these advances, recent work has started to learn related algebraic choices such as the metric used by Clifford group equivariant networks (Ali et al., 2024). Specifically, the fundamental question of whether the grade or type of an observed feature channel can be inferred directly from transformation evidence has not yet been isolated as a distinct empirical and theoretical problem.

The gap matters because different geometric types can be observationally indistinguishable under some transformation families and distinguishable under others. For example, in three dimensions a scalar and a pseudoscalar both remain unchanged under proper rotations, whereas a reflection leaves the scalar unchanged and flips the pseudoscalar sign. Similarly, when bivectors are written in a Hodge-dual coordinate basis, meaning that they are represented as three-dimensional axial vectors or pseudovectors, vectors and bivectors transform in the same way under proper rotations, whereas they differ by a determinant factor under reflections. Consequently, a model trained only on rotations has no evidence that can separate these pairs, while a model trained with reflections may have enough evidence to do so. This observation suggests that grade assignment should not only be a design choice, and it should also be a measurable property of the data and the transformations that the data expose.

To make this property measurable in practice, we must move beyond the abstract mathematical fact that identical representations are indistinguishable. Therefore, our primary contribution is a practical grade discovery diagnostic that applies this fact to Clifford valued features, where parity sensitive ambiguities can silently persist under common rotation only training or validation protocols.

Research question. Given paired observations of a feature channel before and after known geometric transformations, when is its Clifford grade identifiable from the data and when is the grade assignment only a modelling assumption?

We study this issue through a focused diagnostic problem called grade discovery. In this work, grade and geometric type refer to the same modelling choice, namely the Clifford algebra category whose coordinates transform as a scalar, vector, bivector or pseudoscalar. The input is a set of feature channels observed before and after known transformations where each channel has a coordinate dimension, and the task is to assign a soft weight over candidate Clifford grades with that same dimension. That said, for each channel the diagnostic forms a soft mixture of candidate representation matrices and fits the mixture by minimizing an equivariant prediction loss. This construction makes the analysis transparent because in the noiseless population setting, the loss has a unique optimum exactly when the candidate representation laws are separated by the observed transformations and the data covariance is non-degenerate.

The paper makes four contributions. First, it formulates grade discovery as a dimension matched type estimation problem for Clifford valued data with the type variables optimized by a differentiable equivariant prediction loss. Second, it proves an identifiability result that explains both success and failure cases. Specifically, rotation only evidence cannot distinguish scalar from pseudoscalar or vector from bivector in the three-dimensional setting considered here, whereas rotations with reflections can distinguish both pairs under mild data conditions. Third, it introduces controlled synthetic experiments that make these claims directly testable and reproducible including a reflection frequency study, a loss landscape visualization, a projected transformation noise stress test and a covariance degeneracy stress test. Fourth, it clarifies how soft type weights can be used in practice through a gradient-based gate, a minimal trainable prediction module, an equivariant prediction comparison against ground-truth typing and an out of distribution transformation ablation.

The scope is intentionally controlled. We do not claim that grade discovery solves type assignment for every Clifford algebra, every GA model or every dataset. Instead, we show that grade assignment can be audited as a practical diagnostic. At the same time, the identifiability argument itself is not tied to the non-degenerate Cl_3 metric (as discussed in Remark 1), the same separation principle applies to degenerate settings such as PGA once the relevant candidate coordinate actions are specified. When a Clifford valued model depends on a specific grade assignment the underlying data should be examined for supporting transformation evidence, and although observational datasets rarely contain exact transformation pairs natively, such pairs are routinely synthesized through geometric data augmentation or equivariance debugging protocols. In these standard workflows, the exact transformation matrices denoted by Q below are explicitly known allowing our diagnostic to be applied without requiring new data collection. The code release regenerates all tables, figures and summary statistics from fixed seeds.

2 Related work

Equivariant neural networks encode transformation laws in the architecture which improves sample efficiency and reliability when the encoded symmetry matches the task (Cohen & Welling, 2016; Bronstein et al., 2021) and this architectural encoding is implemented in steerable convolutional networks and Euclidean equivariance libraries which treat feature fields as typed objects with prescribed group representations (Weiler & Cesa, 2019; Geiger & Smidt, 2022). Although this representational framework informs our setting, our primary objective is not to design a novel equivariant layer. Instead, we ask whether the transformation law of an observed feature channel can be estimated from paired transformed observations.

Clifford algebra models provide an exceptionally compact language for geometric quantities because multivectors inherently consolidate scalars, vectors, bivectors and higher order objects into graded representations (Hestenes & Sobczyk, 1984; Doran & Lasenby, 2003). This mathematical foundation has driven a recent ecosystem of specialized architectures. Notably, Clifford Neural Layers exploit this specific structure to replace real valued convolutions and Fourier operations with Clifford versions for partial differential equation surrogates (Brandstetter et al., 2023). Similarly, GCANs and CGENNs integrate geometric algebra to build learnable group actions and preserve the geometric product (Ruhe et al., 2023b;a), while GATr leverages PGA to process points, directions, translations and rotations simultaneously within a unified transformer framework (Brehmer et al., 2023). These works demonstrate that typed multivectors are useful, while our work studies how evidence for this geometric typing can be inferred directly from the data.

Several neighbouring ideas learn geometric structure rather than predefining it entirely. For example, metric learning for CGENNs makes the underlying metric data-driven while preserving the Clifford equivariant construction (Ali et al., 2024), just as $SE(3)$-equivariant attention chooses representation channels that transform predictably under rotations and translations (Fuchs et al., 2020). Despite these flexibilities, these approaches still assume that the relevant representation content has been selected by the practitioner. Grade discovery complements these existing approaches by treating a small but consequential part of that selection as an estimable quantity.

3 Problem Setup and Method

3.1 Geometric Notation and Candidate Types

Let the ambient vector space be three-dimensional Euclidean space. We write this space as V, and we identify it with $\mathbb{R}^3$ after choosing an orthonormal basis e_1, e_2 and e_3. The real Clifford algebra generated by this basis is written as Cl_3 and its product satisfies the orthonormal relation for basis vectors. The symbol δ_{ij} denotes the Kronecker delta, which equals one when the indices i and j are equal and zero otherwise. The defining relation is:

$$e_i e_j + e_j e_i = 2\delta_{ij} \tag{1}$$

A multivector in Cl_3 decomposes into grades zero, one, two and three. Grade zero is a scalar, grade one is a vector, grade two is a bivector and grade three is a pseudoscalar. We use the Hodge-dual bivector basis $B_1 = e_2e_3$, $B_2 = e_3e_1$ and $B_3 = e_1e_2$, because this basis represents bivectors as three-dimensional axial-vector coordinates and makes the comparison between vectors and bivectors transparent. We focus on $O(3)$, the group of rotations and reflections in three dimensions, rather than $E(3)$, the Euclidean group that also includes translations, because the representation laws studied here describe quantities that are unchanged by translation and capture magnitude, direction, oriented area or orientation sign. An orthogonal transformation is represented by a matrix $Q \in O(3)$ and its determinant is written as $\det(Q)$. The four representation matrices used in this work are:

$$\begin{aligned} \rho_{\mathrm{scal}}(Q) &= 1 \\ \rho_{\mathrm{pscal}}(Q) &= \det(Q) \\ \rho_{\mathrm{vec}}(Q) &= Q \\ \rho_{\mathrm{biv}}(Q) &= \det(Q)Q \end{aligned} \tag{2}$$

Equation (2) is derived in Appendix A and it implies that the scalar and pseudoscalar representations are identical on the rotation group $SO(3)$, where $\det(Q) = 1$. It also implies that the vector and bivector representations are identical on $SO(3)$ in the chosen Hodge-dual bivector basis. However, both pairs become distinguishable when reflections are observed.

The grade discovery problem is dimension matched. For instance, a one-dimensional feature channel can be typed as a scalar or a pseudoscalar, while a three-dimensional feature channel can be typed as a vector or a bivector. This restriction avoids comparing representation matrices of different sizes and isolates the ambiguity that appears even when the coordinate dimension is already known. We define the candidate sets for one-dimensional and three-dimensional channels as:

$$\begin{aligned} \mathcal{C}_1 &= \{\text{scal}\} \cup \{\text{pscal}\} \\ \mathcal{C}_3 &= \{\text{vec}\} \cup \{\text{biv}\} \end{aligned} \tag{3}$$

3.2 Grade Discovery Loss

Consider a feature channel indexed by c. Its coordinate dimension is written as d_c and its candidate type set is $\mathcal{C}_{d_c}$. The observed untransformed examples are $x_{cj} \in \mathbb{R}^{d_c}$ for item index j, and the transformed observations are $y_{csj} \in \mathbb{R}^{d_c}$ where s indexes a known orthogonal matrix Q_s. For each candidate type $t \in \mathcal{C}_{d_c}$, the representation matrix is $\rho_t(Q_s)$. The unknown type weight for channel c and candidate t is p_{ct} and all weights for a channel sum to one. The mixed representation used by grade discovery is:

$$R_c^{p_c}(Q_s) = \sum_{t \in \mathcal{C}_{d_c}} p_{ct} \rho_t(Q_s) \tag{4}$$

The finite sample loss for one channel is the mean squared equivariant prediction error:

$$\ell_c(p_c) = \frac{1}{SN_c} \sum_{s=1}^{S} \sum_{j=1}^{N_c} \| y_{csj} - R_c^{p_c}(Q_s) x_{cj} \|_2^2 \tag{5}$$

The full discovery objective sums the channel losses. We use logits α_{ct} when the soft weights are learned by gradient descent. The softmax parameterization is:

$$p_{ct} = \frac{\exp(\alpha_{ct})}{\sum_{u \in \mathcal{C}_{d_c}} \exp(\alpha_{cu})} \tag{6}$$

When embedded as a trainable soft gate, this weight can optionally be hardened via an argmax following an initial continuous training phase.

To check that this diagnostic also behaves sensibly when optimized together with another trainable parameter, we evaluate a minimal prediction module in the experiments. The scalar gain for channel c is denoted by γ_c and the module predicts the transformed observation as:

$$\widehat{y}_{csj} = \gamma_c R_c^{p_c}(Q_s) x_{cj} \tag{7}$$

This module is not intended to be a full Clifford neural architecture. Instead, it isolates the practical question of whether the soft gate remains aligned with the closed-form diagnostic when the gate is trained by gradient descent inside a differentiable prediction layer.

For the two candidate experiments in this work the same loss has a closed-form least-squares minimizer and this estimator is used for most reported figures because it removes optimization noise. The differentiable version in Eq. (6) is equivalent in purpose and can be embedded in a neural model. Appendix B derives the closed form and Appendix C derives the gradient used by the softmax version.

3.3 Identifiability Result

The central question is whether the loss in Eq. (5) has enough information to identify the true type. The following theorem states the answer for a dimension matched candidate set under noiseless population data based on a group distribution over transformations, a true type $t_\star$ and a random feature vector x with covariance matrix Σ. The matrix Σ is assumed to be positive definite so that the data vary in every coordinate direction.

Theorem 1 (Identifiability from Representation Separation)**.** *Let $\mathcal{C}_d$ be a finite set of candidate types with the same coordinate dimension d. Let Q be a random orthogonal matrix drawn from the observed transformation distribution and let $x \in \mathbb{R}^d$ be a random feature vector with covariance matrix Σ that is positive definite. Suppose the noiseless transformed observation is $y = \rho_{t_\star}(Q)x$. If the representation functions ρ_t are separated in the sense that the only weight vector p satisfying $\sum_{t\in\mathcal{C}_d} p_t \rho_t(Q) = \rho_{t_\star}(Q)$ almost surely is the one-hot vector on $t_\star$, then the population grade discovery loss has the unique minimizer at the true type. Conversely, if two candidate representation functions are equal almost surely on the observed transformations, then the population loss cannot distinguish those two types.*

Remark 1 (Metric Agnostic Extension)**.** *The theorem is stated for the $O(3)$ representations used in the experiments, but the proof only uses the induced coordinate matrices $\rho_t(Q)$ and the covariance of the observed feature coordinates. It does not require the Clifford metric itself to be non-degenerate. Consequently, the same separation principle applies to other Clifford or geometric algebra representations, including degenerate PGA such as $\mathrm{Cl}_{3,0,1}$, once each candidate object type is represented by a concrete coordinate action $\rho_t(g)$ under the relevant transformation g. The metric signature affects which representation laws are available and which candidates may coincide on the observed transformation support, but it does not change the diagnostic logic, i.e. a type is identifiable only when its representation law is separated from the alternatives on the observed transformations. In degenerate algebras, homogeneous constraints and ideal components may create null or constrained directions, so the covariance condition should be interpreted on the data supported active coordinate subspace rather than as a global full-rank condition on all algebra coordinates.*

The proof is provided in Appendix D. Because the loss reaches zero exclusively when the mixed and true representations align on the observed transformation support, identifiability is determined by the observed transformation family in the noiseless population setting. In particular, the diagnostic succeeds when the family separates the candidates and fails when the family renders them mathematically identical.

Proposition 1 (Three-Dimensional Grade Ambiguity)**.** *For the representations in Eq.* (2)*, rotation only evidence does not identify scalar versus pseudoscalar or vector versus bivector. If reflections occur with positive probability and the data covariance is positive definite, then these two dimension matched pairs are identifiable.*

The proposition follows directly from $\det(Q) = 1$ for rotations and $\det(Q) = -1$ for reflections. Appendix D gives the full algebraic calculation and building on this result, we can predict how reflection frequency affects recovery in the noiseless two candidate setting. Let r be the probability that an observed transformation is a reflection, let S be the number of observed transformations and let $W_{\mathrm{true}}(r, S)$ be the expected soft weight assigned to the true type under the rule that no reflection gives the uninformative weight one half while at least one reflection identifies the type. The resulting expression is:

$$W_{\mathrm{true}}(r, S) = 1 - \frac{1}{2}(1 - r)^S \tag{8}$$

Appendix E derives Eq. (8) step by step. This theoretical calculation directly motivates the subsequent reflection frequency experiments because it gives a quantitative estimate of how rare parity revealing transformations can be before grade recovery becomes fundamentally unreliable.

4 Controlled Benchmark

4.1 Data Generation and Evaluation

The benchmark contains four channel types. The one-dimensional channels are scalar and pseudoscalar, while the three-dimensional channels are vector and bivector. For each channel, untransformed examples are sampled from a standard normal distribution and a transformed observation is generated by applying the true representation matrix from Eq. (2) and adding independent Gaussian noise. The default noise standard deviation is 0.02 which is chosen to simulate mild measurement or numerical precision errors without completely destroying the geometric signal (we systematically evaluate higher noise regimes up to 0.5 in Section 4.2).

We compare rotation only data, rotations with reflections and training distributions where the reflection probability varies from zero to one half. Unless otherwise stated reported values are means and standard errors over ten random seeds, while the reflection frequency curve uses twenty seeds because rare reflections create higher finite sample variability. We also include two stress tests. The first generates transformed observations from exact matrices but gives the estimator matrices projected from noisy measurements back onto the same component of $O(3)$ that tests sensitivity to inaccurate but still orthogonal transformation matrices. The second uses a generic two law diagnostic pair that differs only along a low variance coordinate which makes the covariance assumption in the theorem operational. The exact matrices for this diagnostic pair are provided in Appendix G. The grade discovery estimator observes the transformation matrix and the paired examples but it does not observe the true type label, whereas existing Clifford valued models typically require the grade of each feature channel to be specified before training, so there is no established algorithmic baseline for grade discovery and as such we compare learned typing with three reference baselines. The ground-truth typing baseline uses the generator grade and the complementary grade baseline assigns each channel to the other dimension matched candidate, that is to the Hodge-dual grade $3 - k$ for a grade-k channel in Cl_3 with scalar $\leftrightarrow$ pseudoscalar and vector $\leftrightarrow$ bivector, thereby measuring the cost of specifying the incorrect parity behaviour under reflections. The identity baseline ignores the transformation.

4.2 Results

Table 1 shows the central identifiability result together with the evidence denominator from the closed-form estimator. Under rotation only evidence, every channel receives weight one half on the true type. This is not a training failure, because it is the predicted outcome when the evidence denominator in the two candidate least-squares estimator is zero. When reflections are added, the estimator assigns weight nearly one to the true type in the displayed run for all four channels. The denominator is the sum of squared differences between the two candidate predictions, i.e. a zero denominator means that the candidates make identical predictions on the observed transformations and examples and a positive denominator means that the observations contain evidence that can separate the candidates.

Table 1: Single run type recovery and evidence denominator. The mixture mean squared error reflects the noise level because the fitted representation matches the true type when identification is possible.

Transformation family	True type	True type weight	Mixture MSE	Evidence denominator
rotations	scalar	0.500	0.0004	0.0
rotations	pseudoscalar	0.500	0.0004	0.0
rotations	vector	0.500	0.0004	0.0
rotations	bivector	0.500	0.0004	0.0
rotations + reflections	scalar	1.000	0.0004	8696.5
rotations + reflections	pseudoscalar	1.000	0.0004	9302.4
rotations + reflections	vector	1.000	0.0004	28261.1
rotations + reflections	bivector	1.000	0.0004	26741.4

The left panel of Figure 1 reports the effect of varying the number of observed transformations with means and standard errors computed over ten random seeds. Under rotation only evidence, the weight assigned

to the true type remains at 1/2 for every sample size reflecting the non-identifiability predicted by the representation laws. When rotations and reflections are both available, one observed transformation gives a mean true type weight of 0.712 with standard error 0.046 because some seeds still contain no reflection. Four observed transformations raise the mean weight to 0.950 with standard error 0.028 and eight or more transformations make the mean weight exceed 0.999. Consequently, the estimator behaves as a diagnostic for whether the training distribution exposes the parity revealing information required to support the type choice.

The right panel of Figure 1 fixes the number of sampled transformations at $S = 16$ and varies the probability that each sampled transformation Q_s is a reflection. The simulation closely follows the noiseless presence calculation in Eq. (8). With no reflections, the weight assigned to the true type is exactly 1/2 and with a reflection probability of 1% the mean weight assigned to the true type is 0.581 with standard error 0.016. Increasing the reflection probability to 10% raises the mean to 0.906 with standard error 0.018, and at 20% or higher the mean exceeds 0.999 in this benchmark. Thus, grade discovery does not require a balanced rotation-reflection training distribution, but extremely rare reflections leave many finite samples ambiguous.

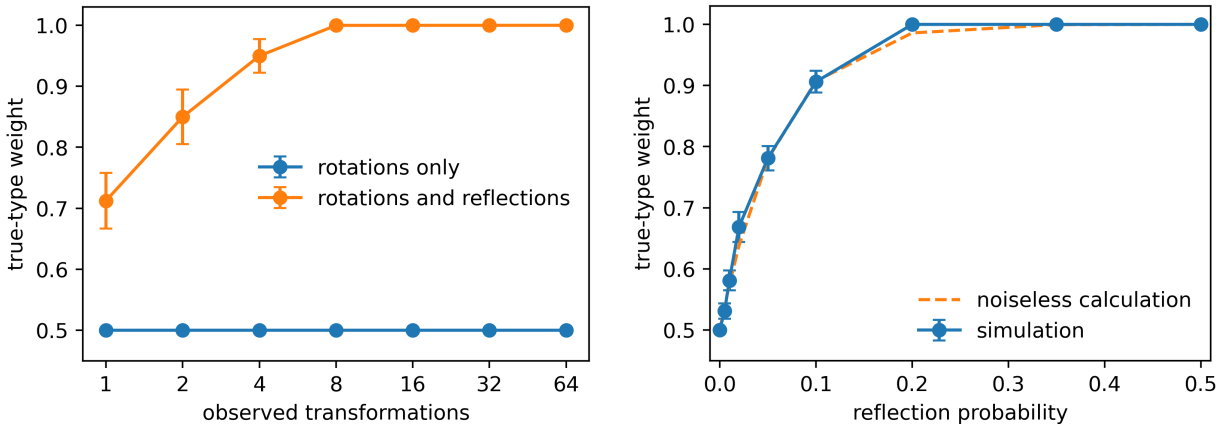

Figure 1: Data requirements for grade discovery. The left panel shows mean weight assigned to the true type as the number of observed transformations increases, with error bars showing standard error over ten seeds. The right panel shows mean weight assigned to the true type as the reflection probability per observed transformation varies, with error bars showing standard error over twenty seeds and the dashed curve showing the noiseless probability of seeing at least one parity revealing transformation.

Figure 2 visualizes the loss geometry behind the reflection frequency result. The plot uses an unconstrained two weight version of the scalar-pseudoscalar loss with the horizontal axis representing the scalar weight and the vertical axis representing the pseudoscalar weight. Under rotation only evidence, the two candidates act identically, so every point on the weight simplex gives the same zero loss. Adding reflections breaks this flat valley and creates a unique minimum at the true scalar solution (Appendix F).

The left panel of Figure 3 compares the differentiable softmax gate with the closed-form solution on the bivector channel, where it is observed that the gradient-based weight starts at 1/2 because the two logits are initialized equally. As the loss is optimized, the weight moves toward the closed-form value reaching a mean true type weight of 0.978 after 160 gradient steps. The closed-form solution is useful for analysis and diagnostics because it removes optimizer effects, while the softmax trajectory shows how the same estimator behaves when implemented as a trainable differentiable gate.

The right panel of Figure 3 varies the observation noise standard deviation. Under rotation only evidence, the weight assigned to the true type remains at 1/2 because the candidate predictions are identical on all observed transformations. Under rotations and reflections, the weight remains near one across the tested noise range and reaches a mean value of 0.997 at noise standard deviation 0.5. This experiment does not imply that the method is robust to arbitrary noise, but rather it shows that in this simple finite sample

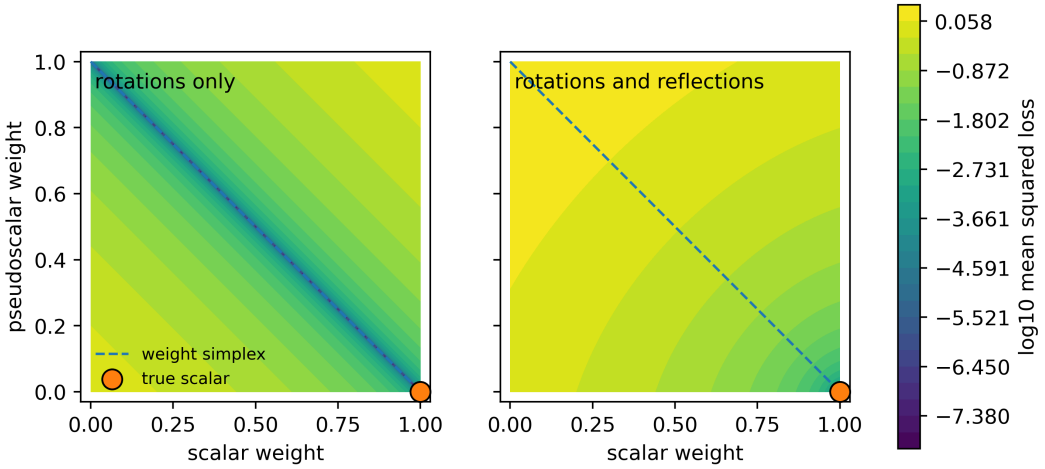

Figure 2: Loss landscape for scalar versus pseudoscalar weights in the noiseless setting. The dashed diagonal is the weight simplex and the orange marker with black outline indicates the true scalar solution. Rotation only evidence creates a flat valley along the simplex, while rotations with reflections produce a unique minimum at the true scalar type.

setting once reflections separate the candidate representation laws, the least-squares estimator continues to assign high weight to the correct type across the tested noise levels.

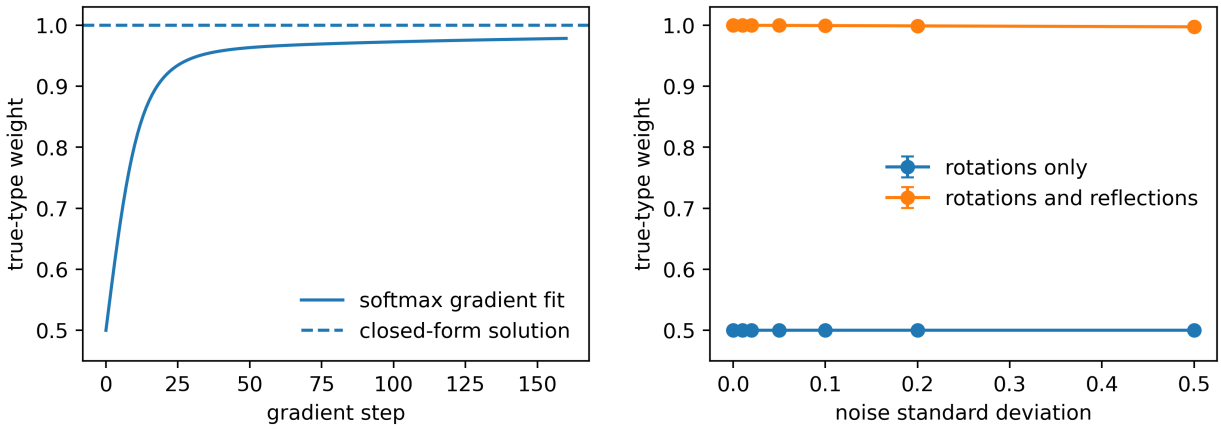

Figure 3: Optimization and noise behaviour. The left panel reports the mean weight assigned to the true type over five seeds for the differentiable two candidate softmax gate, with the dashed line showing the closed-form solution for the same datasets. The right panel reports mean weight assigned to the true type as Gaussian observation noise increases, with error bars showing standard error over ten seeds.

Figure 4 stress tests two assumptions behind the diagnostic. The left panel generates training and test observations from exact transformations but gives the estimator approximate matrices that are projected back to the same connected component of $O(3)$. Specifically, for Gaussian matrix noise E_s and perturbation scale η, the supplied matrix is:

$$\widetilde{Q}_s = \mathrm{Proj}_{O(3),\det(Q_s)}(Q_s + \eta E_s) \tag{9}$$

Here, $\mathrm{Proj}_{O(3),\det(Q_s)}$ denotes the nearest orthogonal matrix with the same determinant sign as Q_s, computed by singular value decomposition. Consequently, the experiment tests inaccurate but still orthogonal transformation matrices rather than non-orthogonal representation law misspecification. As the perturbation scale increases from zero to 0.35, we plot the mean true type weight and test mean squared error to

show how this assumption degrades. The right panel studies the covariance condition using a generic two law diagnostic pair whose candidates differ only on the third coordinate. This stress test is not an additional Clifford grade claim. Rather, it shows why low variance in the coordinate where candidate laws differ leads to a smaller evidence denominator and therefore less stable type recovery under observation noise.

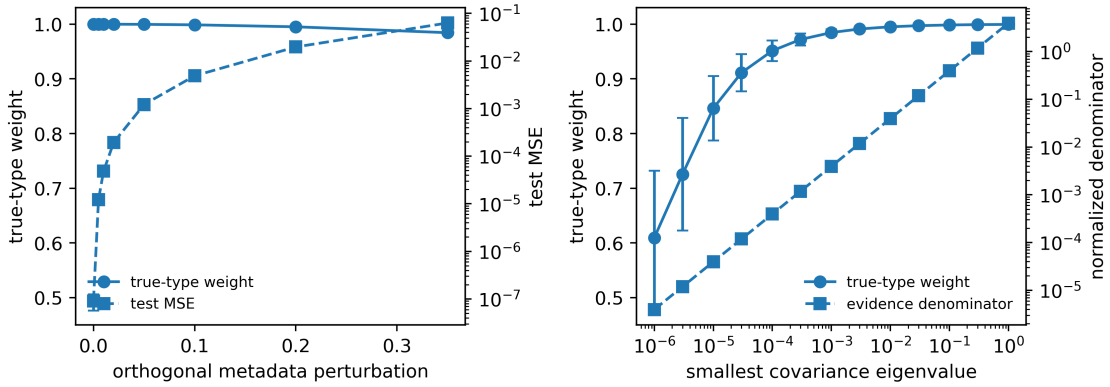

Figure 4: Diagnostic stress tests. The left panel reports true type weight and test mean squared error when the data are generated from exact transformations but the estimator receives noisy matrices projected back to the same component of $O(3)$. The right panel reports true type weight and the normalized evidence denominator in a theorem-assumption stress test where the smallest covariance eigenvalue controls the amount of variation in the coordinate that separates two generic laws.

Table 2 reports the minimal trainable prediction module from Eq. (7). When the module is trained and tested on rotation only data, it reaches nearly zero test error while keeping the type weight at one half which confirms that rotations do not provide parity information. However, the same rotation only trained module fails on reflection containing tests because the learned mixture has not identified the parity behaviour. When reflections are included during training, the soft gate moves toward the correct type and the test error on rotations and reflections becomes small. This experiment is intentionally minimal and as such it is not evidence of performance in full Clifford neural networks and it only checks that the diagnostic gate can be optimized jointly with another parameter in a small prediction layer.

Table 2: Minimal trainable prediction module with a soft grade gate and a learned scalar gain. The table reports training transformations, test transformations, final true type weight, learned gain, test mean squared error and the number of seeds.

Train transformations	Test transformations	Final true type weight	Learned γ	Test MSE	Seeds
rotations only	rotations only	0.500 ± 0.000	1.000 ± 0.000	1.41e-07 ± 5.73e-08	10
rotations only	rotations + reflections	0.500 ± 0.000	1.000 ± 0.000	5.11e-01 ± 1.37e-02	10
rotations + reflections	rotations + reflections	0.929 ± 0.001	1.071 ± 0.005	6.23e-03 ± 3.01e-04	10

The equivariant prediction experiment in Table 3 and Figure 5 tests whether the discovered type matters downstream. The ground-truth typing baseline has zero error on noiseless test transformations. Learned typing reaches mean squared error 4.07×10^{-8} with standard error 1.19×10^{-8} which is numerically indistinguishable from the ground-truth typing baseline at the plotted scale. In contrast, the complementary grade baseline has mean squared error 1.95 with standard error 0.081, because reflections flip the parity of pseudoscalars and bivectors but not scalars and vectors. The identity baseline also fails because it ignores rotations and reflections entirely.

Table 4 clarifies the behaviour of learned typing when the test transformations differ from the training transformations. Training on rotation only data and testing on rotations gives almost zero equivariant

Table 3: Equivariant prediction error on unseen rotations and reflections after fitting type weights on noisy training pairs. Values are means and standard errors over ten seeds.

Method	Mean equivariant prediction MSE	Standard error
ground-truth typing	0.00e+00	0.00e+00
learned typing	4.07e-08	1.19e-08
complementary grade	1.95e+00	8.10e-02
identity transform	1.47e+00	4.15e-02

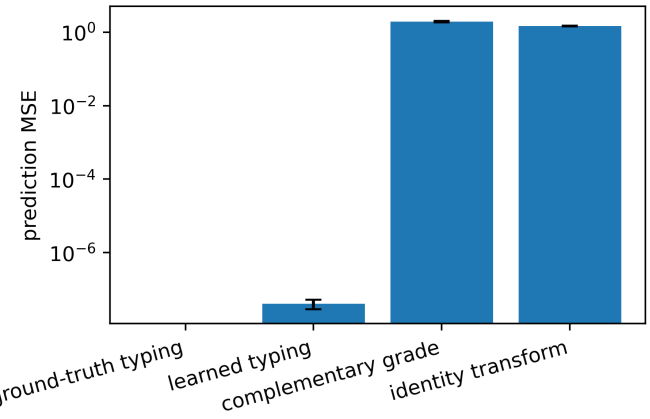

Figure 5: Equivariant prediction mean squared error on a logarithmic scale. Learned typing matches ground-truth typing and avoids the parity error of the complementary grade type.

Table 4: Out of distribution transformation ablation. The table separates uncertainty about the type from mis-specification of the representation law used at test time. Values are means and standard errors over ten seeds.

Setting	True type weight	Learned equivariant MSE
train rotations only, test rotations	0.500 ± 0.000	4.91e-32 ± 1.38e-33
train rotations only, test rotations and reflections	0.500 ± 0.000	5.14e-01 ± 1.56e-02
train rotations and reflections, test rotations and reflections	1.000 ± 0.000	7.54e-08 ± 3.71e-08
train rotations and reflections, test scaling and shear with supplied general linear laws	1.000 ± 0.000	3.28e-09 ± 1.42e-09
train rotations and reflections, test scaling and shear with orthogonal only laws	1.000 ± 0.000	4.28e-02 ± 1.91e-03

prediction error, even though the type remains unidentified because the dimension matched candidates make identical predictions on rotations. However, training on rotation only data and testing on rotations with reflections gives a learned equivariant prediction error of 0.514 because the fitted 1/2-1/2 mixture does not match the required parity behaviour on reflection examples. On the other hand, training with reflections and testing on rotations with reflections gives near zero error. The same learned types also transfer to scaling and shear when evaluation uses the correct general linear representation laws for those transformations. This row should not be read as evidence that grade discovery inferred scaling or shear laws from data, because those non-orthogonal representation laws are supplied externally at evaluation time. Thus this experiment tests transfer of an already identified grade under supplied representation laws, not discovery of new non-orthogonal laws. The final row shows the boundary of this conclusion, so if scaling and shear examples

are evaluated with orthogonal only representation laws, the learned type can still be confident while the transformation law itself is misspecified producing nonzero error.

5 Discussion and Limitations

The results support a specific claim. The grade discovery diagnostic can recover dimension matched Clifford types when the observed transformations distinguish their representation laws and it cannot recover those types when the transformation family makes the laws identical (this claim is useful because it turns a hidden modelling assumption into a testable property).

The soft type weights should be handled according to the available evidence. When the maximum weight is close to one and the validation loss supports the same conclusion, hardening the type with an argmax produces a clean Clifford representation and removes unnecessary mixture routing. When the weights remain near one half, hardening is not supported by the data. In that case, the output of grade discovery is the ambiguity itself and it becomes necessary to either collect transformations that distinguish the candidates or state the type choice as a prior assumption. To clarify when this occurs Table 5 summarizes the primary failure modes of the diagnostic.

Table 5: Failure mode summary for the grade discovery diagnostic.

Failure mode	Diagnostic interpretation
No parity revealing transformations	The candidate grades are unidentifiable from the observed evidence
Rare reflections	Finite samples may contain no separating transformation and remain ambiguous
Degenerate or low variance evidence	Candidate laws may differ in directions where the data provide little signal
Wrong representation law	The diagnostic may be confident about type while extrapolation fails under a misspecified transformation law
Approximate orthogonal matrices	Perturbing the supplied matrices and projecting them back to $O(3)$ can reduce true type weight and increase prediction error

Several limitations define the boundary of the present work. First, the experiments are controlled and synthetic by design which makes the identifiability claims directly testable but does not establish performance on large real datasets. Second, the estimator assumes paired observations and either exact or sufficiently accurate transformation matrices. The projected transformation noise stress test shows that this assumption is measurable rather than harmless. Third, the current formulation compares only candidates with the same coordinate dimension which is appropriate for scalar versus pseudoscalar and vector versus bivector ambiguity, but it does not decide whether a raw channel should be one-dimensional or three-dimensional. Fourth, the method selects among predefined representation laws. As the scaling and shear ablation shows, confident type discovery does not repair a wrong transformation law outside the modelled family. Finally, the paper studies the representation law of feature channels and a minimal trainable prediction module rather than a full Clifford neural architecture. These boundaries keep the diagnostic isolated before it is combined with more complex models.

A natural next step is to attach the softmax version of the estimator to a Clifford neural network and train task loss and grade discovery loss jointly. Another direction is to extend the candidate set to conformal and projective GA objects such as points, lines, planes and motors. The identifiability principle should remain the same, although the representation spaces and degeneracies will be richer.

6 Conclusion

This work studies grade discovery for Clifford valued features where the main finding is that the geometric type of a feature channel is learnable only to the extent that the observed transformations separate the candidate representation laws. In three dimensions, rotation only evidence leaves scalar versus pseudoscalar and vector versus bivector ambiguous, while reflections resolve both ambiguities under non-degenerate data. The least-squares estimator, the differentiable softmax version, the trainable gate experiment and the stress tests provide a reproducible grade discovery diagnostic for checking whether a Clifford valued feature assignment is supported by transformation evidence.

References

Riccardo Ali, Paulina Kulytė, Haitz Sáez de Ocáriz Borde, and Pietro Lio. Metric learning for clifford group equivariant neural networks. *arXiv preprint arXiv:2407.09926*, 2024.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K. Gupta. Clifford neural layers for pde modeling. In *International Conference on Learning Representations*, 2023.

Johann Brehmer, Pim De Haan, Sönke Behrends, and Taco S Cohen. Geometric algebra transformer. *Advances in Neural Information Processing Systems*, 36:35472–35496, 2023.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999. PMLR, 2016.

Chris Doran and Anthony Lasenby. *Geometric Algebra for Physicists*. Cambridge University Press, 2003.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

David Hestenes and Garret Sobczyk. *Clifford Algebra to Geometric Calculus*. D. Reidel Publishing Company, 1984.

David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks. *Advances in neural information processing systems*, 36:62922–62990, 2023a.

David Ruhe, Jayesh K. Gupta, Steven de Keninck, Max Welling, and Johannes Brandstetter. Geometric clifford algebra networks. In *International Conference on Machine Learning*. PMLR, 2023b.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

A Derivation of the Grade Representations

We outline the derivation of Eq. (2) below. Let $Q \in O(3)$ be an orthogonal matrix with entries q_{ai}, where a indexes the output basis coordinate and i indexes the input basis coordinate. The transformed basis vector takes the form:

$$Qe_i = \sum_{a=1}^{3} q_{ai} e_a \tag{10}$$

For a vector $v = \sum_i v_i e_i$ substitution gives:

$$Qv = Q\left(\sum_{i=1}^{3} v_i e_i\right) \tag{11}$$

$$= \sum_{i=1}^{3} v_i Q e_i \tag{12}$$

$$= \sum_{i=1}^{3} v_i \sum_{a=1}^{3} q_{ai} e_a \tag{13}$$

$$= \sum_{a=1}^{3} \left(\sum_{i=1}^{3} q_{ai} v_i\right) e_a \tag{14}$$

and as such the coordinate vector transforms as:

$$v \mapsto Qv \tag{15}$$

A scalar has no orientation and is unchanged by Q, so its representation is:

$$\rho_{\text{scal}}(Q) = 1 \tag{16}$$

Let $I = e_1 e_2 e_3$ be the unit pseudoscalar. Applying Q to the oriented volume element gives the determinant factor:

$$QI = (Qe_1)(Qe_2)(Qe_3) \tag{17}$$

$$= \det(Q) e_1 e_2 e_3 \tag{18}$$

$$= \det(Q) I \tag{19}$$

Thus the pseudoscalar coordinate transforms as:

$$\rho_{\text{pscal}}(Q) = \det(Q) \tag{20}$$

The Hodge-dual bivector basis used in the paper is $B_1 = e_2 e_3$, $B_2 = e_3 e_1$, and $B_3 = e_1 e_2$. A bivector with axial coordinates $b \in \mathbb{R}^3$ can be written as the dual of a vector coordinate representation, and under an orthogonal transformation the pseudoscalar contributes $\det(Q)$ and the axial coordinates rotate by Q which gives:

$$b \mapsto \det(Q) Q b \tag{21}$$

Finally combining these four laws gives:

$$\begin{aligned} \rho_{\text{scal}}(Q) &= 1 \\ \rho_{\text{pscal}}(Q) &= \det(Q) \\ \rho_{\text{vec}}(Q) &= Q \\ \rho_{\text{biv}}(Q) &= \det(Q) Q \end{aligned} \tag{22}$$

B Least-Squares Form of the Grade Discovery Loss

This appendix derives the closed-form estimator used for the two candidate experiments. Let the two candidate representation matrices for transformation Q_s be A_s and B_s. The scalar mixture weight on A_s is p, and the weight on B_s is $1 - p$. Let us define the difference matrix as:

$$D_s = A_s - B_s \tag{23}$$

The mixed prediction can be written as:

$$\widehat{y}_{sj}(p) = (pA_s + (1-p)B_s)\, x_j \tag{24}$$
$$= (B_s + p(A_s - B_s))\, x_j \tag{25}$$
$$= B_s x_j + pD_s x_j \tag{26}$$

The loss is:

$$L(p) = \frac{1}{SN} \sum_{s=1}^{S} \sum_{j=1}^{N} \|y_{sj} - B_s x_j - pD_s x_j\|_2^2 \tag{27}$$

Let $r_{sj} = y_{sj} - B_s x_j$ and $z_{sj} = D_s x_j$. Expanding the squared norm gives:

$$L(p) = \frac{1}{SN} \sum_{s,j} \|r_{sj} - pz_{sj}\|_2^2 \tag{28}$$
$$= \frac{1}{SN} \sum_{s,j} \left(r_{sj}^\top r_{sj} - 2p r_{sj}^\top z_{sj} + p^2 z_{sj}^\top z_{sj} \right) \tag{29}$$

The derivative with respect to p is:

$$\frac{dL}{dp} = \frac{1}{SN} \sum_{s,j} \left(-2 r_{sj}^\top z_{sj} + 2p z_{sj}^\top z_{sj} \right) \tag{30}$$

Setting the derivative to zero gives:

$$0 = \sum_{s,j} \left(-r_{sj}^\top z_{sj} + p z_{sj}^\top z_{sj} \right) \tag{31}$$
$$p \sum_{s,j} z_{sj}^\top z_{sj} = \sum_{s,j} r_{sj}^\top z_{sj} \tag{32}$$
$$p = \frac{\sum_{s,j} r_{sj}^\top z_{sj}}{\sum_{s,j} z_{sj}^\top z_{sj}} \tag{33}$$

Because p is a soft weight, the implemented estimator clips this value to the interval $[0, 1]$. If the denominator is zero, the two candidate predictions are identical on the observed data and the implementation returns $p = 1/2$.

C Gradient for the Differentiable Gate

The differentiable gate uses logits α_t and softmax weights p_t. The softmax is:

$$p_t = \frac{\exp(\alpha_t)}{\sum_{u \in \mathcal{C}} \exp(\alpha_u)} \tag{34}$$

The derivative of the softmax is:

$$\frac{\partial p_t}{\partial \alpha_u} = p_t \left(\mathbf{1}\{t = u\} - p_u \right) \tag{35}$$

Let the prediction for transformation s and item j be $\widehat{y}_{sj} = \sum_t p_t \rho_t(Q_s) x_j$, and let $e_{sj} = \widehat{y}_{sj} - y_{sj}$. The derivative of the loss with respect to p_t is:

$$\frac{\partial L}{\partial p_t} = \frac{2}{SN} \sum_{s=1}^{S} \sum_{j=1}^{N} e_{sj}^\top \rho_t(Q_s) x_j \tag{36}$$

Combining Eq. (35) and Eq. (36) gives the logit gradient:

$$\frac{\partial L}{\partial \alpha_u} = \sum_{t \in \mathcal{C}} \frac{\partial L}{\partial p_t} p_t \left(\mathbf{1}\{t = u\} - p_u \right) \tag{37}$$

D Proof of the Identifiability Theorem

Let the true type be $t_\star$ and define the population residual matrix for a weight vector p as:

$$M_p(Q) = \rho_{t_\star}(Q) - \sum_{t \in \mathcal{C}_d} p_t \rho_t(Q) \tag{38}$$

In the noiseless population setting, $y = \rho_{t_\star}(Q)x$. The population loss is:

$$\mathcal{L}(p) = \mathbb{E}\left[\left\| \rho_{t_\star}(Q)x - \sum_{t \in \mathcal{C}_d} p_t \rho_t(Q)x \right\|_2^2 \right] \tag{39}$$

$$= \mathbb{E}\left[\|M_p(Q)x\|_2^2 \right] \tag{40}$$

Conditioning on Q and using $\Sigma = \mathbb{E}[xx^\top]$ gives:

$$\mathbb{E}_x\left[\|M_p(Q)x\|_2^2 \mid Q \right] = \mathbb{E}_x\left[x^\top M_p(Q)^\top M_p(Q)x \mid Q \right] \tag{41}$$

$$= \operatorname{tr}\left(M_p(Q)^\top M_p(Q)\Sigma \right) \tag{42}$$

If Σ is positive definite, this conditional expectation is zero only when $M_p(Q) = 0$. Therefore $\mathcal{L}(p) = 0$ if and only if the mixed representation equals the true representation almost surely. The separation assumption states that only the one-hot vector on $t_\star$ has this property which proves uniqueness. Conversely, if two candidate representations are equal almost surely on the observed transformations replacing soft weight between those two candidates does not change the mixed representation, so the loss cannot distinguish them.

For the three-dimensional proposition, rotations satisfy $\det(Q) = 1$. Substitution into Eq. (2) gives the rotation only equalities:

$$\rho_{\text{scal}}(Q) = \rho_{\text{pscal}}(Q) = 1 \tag{43}$$

The vector and bivector equality under rotations is:

$$\rho_{\text{vec}}(Q) = \rho_{\text{biv}}(Q) = Q \tag{44}$$

For a reflection, $\det(Q) = -1$. The scalar and pseudoscalar laws become:

$$\rho_{\text{scal}}(Q) = 1 \qquad \rho_{\text{pscal}}(Q) = -1 \tag{45}$$

The vector and bivector laws become:

$$\rho_{\text{vec}}(Q) = Q \qquad \rho_{\text{biv}}(Q) = -Q \tag{46}$$

Thus any reflection separates both candidate pairs, and the positive definite covariance condition ensures that this separation is visible to the loss.

E Reflection Frequency Calculation

Let r be the probability that one sampled transformation is a reflection and let S be the number of sampled transformations. The probability that a single sampled transformation is not a reflection is:

$$1 - r \tag{47}$$

The probability that none of the S sampled transformations is a reflection is:

$$(1 - r)^S \tag{48}$$

Under the idealized rule used in the calculation, no reflection gives true type weight $1/2$, and at least one reflection gives true type weight 1. Therefore the expected true type weight is:

$$W_{\text{true}}(r, S) = \frac{1}{2}(1 - r)^S + 1 - (1 - r)^S \tag{49}$$

$$= 1 - \frac{1}{2}(1 - r)^S \tag{50}$$

F Loss Landscape Calculation

The loss landscape figure uses the scalar-pseudoscalar case with a true scalar. Let w_s be the scalar weight, let w_p be the pseudoscalar weight and let $d = \det(Q)$. The prediction is:

$$\widehat{y} = (w_s + w_p d)x \tag{51}$$

Because the true output is $y = x$, the expected squared loss is:

$$L(w_s, w_p) = \mathbb{E}\left[x^2 \left(1 - w_s - w_p d\right)^2\right] \tag{52}$$

For rotation only evidence, $d = 1$ with probability one, so the loss is:

$$L_{SO(3)}(w_s, w_p) = \mathbb{E}[x^2] \left(1 - w_s - w_p\right)^2 \tag{53}$$

This expression is zero on the line $w_s + w_p = 1$, which is the flat valley shown on the weight simplex. If rotations and reflections occur with equal probability, then $d = 1$ and $d = -1$ each occur with probability $1/2$ giving:

$$L_{O(3)}(w_s, w_p) = \frac{\mathbb{E}[x^2]}{2} \left(1 - w_s - w_p\right)^2 + \frac{\mathbb{E}[x^2]}{2} \left(1 - w_s + w_p\right)^2 \tag{54}$$

$$= \mathbb{E}[x^2] \left((1 - w_s)^2 + w_p^2\right) \tag{55}$$

The unique minimizer is $w_s = 1$ and $w_p = 0$ which corresponds to the true scalar type.

G Robustness Stress Tests

The projected transformation noise stress test generates transformed observations from exact matrices Q_s but gives the estimator projected approximate matrices $\widetilde{Q}_s$. The perturbation model is:

$$\widetilde{Q}_s = \text{Proj}_{O(3), \det(Q_s)}(Q_s + \eta E_s) \tag{56}$$

Here, η is the reported orthogonal matrix perturbation scale, the entries of E_s are independent standard normal random variables and the projection is computed by singular value decomposition while preserving

the determinant sign of Q_s. The data are generated with Q_s, while the diagnostic evaluates candidate laws with $\widetilde{Q}_s$. This directly measures the effect of inaccurate but still orthogonal transformations on true type weight and equivariant prediction error.

The covariance degeneracy stress test is a generic two law diagnostic rather than a new Clifford grade experiment. The two candidate laws are:

$$A = I_3 \qquad B = \mathrm{diag}(1, 1, -1) \tag{57}$$

The feature covariance is:

$$\Sigma_\lambda = \mathrm{diag}(1, 1, \lambda) \tag{58}$$

The two laws differ only in the third coordinate, so the evidence denominator scales with the variance λ in that coordinate. This setup operationalizes the positive definite covariance assumption in Theorem 1 by showing that a very small separating variance leads to weak finite sample evidence under observation noise.

H General Linear Laws used in the Scaling and Shear Ablation

The main theory uses $O(3)$, while the scaling and shear ablation supplies representation laws for an invertible matrix M. Vectors transform as:

$$v \mapsto Mv \tag{59}$$

Pseudoscalars transform by the determinant:

$$I \mapsto \det(M) I \tag{60}$$

Bivectors in axial coordinates transform by the cofactor matrix. The identity is:

$$(Ma) \times (Mb) = \det(M) M^{-\top} (a \times b) \tag{61}$$

Therefore the supplied general linear bivector law is:

$$\rho_{\mathrm{biv}}(M) = \det(M) M^{-\top} \tag{62}$$

This ablation tests whether an already identified grade transfers when the correct non-orthogonal representation law is supplied externally. It does not test whether the diagnostic discovers new non-orthogonal laws.

I Experimental implementation details

Table 6 gives the compact reproducibility details and the code release regenerates every table and figure from fixed seeds.

J Gradient Calculation for the Minimal Trainable Prediction Module

The minimal module predicts $\widehat{y}_{sj} = \gamma \left(pA_s + (1-p)B_s\right) x_j$. Define:

$$m_{sj}(p) = \left(B_s + p(A_s - B_s)\right) x_j \tag{63}$$

The prediction and residual are:

$$\widehat{y}_{sj} = \gamma m_{sj}(p) \qquad e_{sj} = \widehat{y}_{sj} - y_{sj} \tag{64}$$

The derivative with respect to γ is:

$$\frac{\partial L}{\partial \gamma} = \frac{2}{SN} \sum_{s,j} e_{sj}^\top m_{sj}(p) \tag{65}$$

Table 6: Reproducibility details for the controlled benchmark.

Item	Value
Training examples per channel	128 (96 for the sample complexity, reflection frequency and noise robustness sweeps)
Default training transformations	16
Single run table transformations	32
Test examples and transformations	256 examples and 32 transformations
Default observation noise	Gaussian noise with standard deviation 0.02
Observation noise sweep	0, 0.01, 0.02, 0.05, 0.1, 0.2 and 0.5
Reflection probabilities	0, 0.005, 0.01, 0.02, 0.05, 0.10, 0.20, 0.35 and 0.50
Transformation noise values	0, 0.005, 0.01, 0.02, 0.05, 0.10, 0.20 and 0.35
Covariance eigenvalue values	From 10^{-6} to 1 on a logarithmic grid
Seeds	10 seeds unless otherwise stated, 20 for reflection frequency and 5 for the softmax trajectory
Optimizer settings	Adam style scalar updates with 160 steps, learning rate 0.15 for the gate trajectory and 0.05 for the trainable module

Let $D_s = A_s - B_s$. The derivative with respect to p is:

$$\frac{\partial L}{\partial p} = \frac{2\gamma}{SN} \sum_{s,j} e_{sj}^{\top} D_s x_j \tag{66}$$

For the two candidate sigmoid gate $p = \sigma(z)$ the logit derivative is:

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial p} p(1-p) \tag{67}$$