GRASSMLR: RETHINKING MULTINOMIAL LOGISTIC REGRESSION ON GRASSMANN MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Grassmann manifolds represent data as low-dimensional subspaces, effectively capturing the intrinsic geometric structure of high-dimensional data, thus achieving significant success in geometric representation learning. However, existing approaches typically map Riemannian features to Euclidean spaces for final decision, which inevitably introduces semantic distortions and separates the decision process from the underlying geometry, thereby overlooking the geometric power embedded in the Riemannian space. To address this issue, we propose Grassmann Multinomial Logistic Regression **(GrassMLR)**, which performs the classification decision directly on the Grassmann manifold. Specifically, to align classification decisions with the intrinsic geometry of the Grassmann manifold, GrassMLR introduces two learnable Grassmann anchors for each class. For an input subspace, it computes the squared geodesic distances to the two anchors. A distance-difference function is then defined as the difference between these two squared geodesic distances. This formulation induces a decision hypersurface with explicit geometric meaning, thereby ensuring that classification decisions remain consistent with the underlying manifold geometry. Furthermore, to obtain representations better suited for GrassMLR, we propose a Quotient-Space Representation Learning (QSRL) strategy. In particular, we exploit the quotient-space invariance of the Grassmann manifold to apply a learnable Lie-group transformation to subspace representations, preserving the quotient space while obtaining more effective geometric representations. To validate the generalised potential of our design, we apply GrassMLR to multiple downstream tasks. Extensive experiments demonstrate that GrassMLR and QSRL work synergistically to improve the performance of Grassmann neural networks and obtain more discriminative geometric representations. These results validate the effectiveness of geometry-aware decision making and representation optimisation under quotient space invariance.

1 INTRODUCTION

In recent years, Grassmann manifolds have provided a natural framework for characterising the intrinsic geometric structure of data by modelling high-dimensional data as low-dimensional linear subspaces (Edelman et al., 1998; Bendokat et al., 2024; Bronstein et al., 2021). Based on this modelling paradigm, Grassmann manifolds have been widely applied to tasks such as action recognition (Huang et al., 2018; Yu & Xu, 2026), emotion recognition (Yu & Xu, 2026), and EEG signal classification (Wang et al., 2024; Yu & Xu, 2026). In these works, existing methods typically learn subspace representations directly on the Grassmann manifold, thereby obtaining Riemannian features with guaranteed geometric structure (Huang et al., 2018; Wang et al., 2024; Yu & Xu, 2026). However, at the decision stage, these Riemannian features are typically mapped to Euclidean space for decision making (Huang et al., 2018; Wang et al., 2024; Yu & Xu, 2026), as illustrated in Fig. 1. Although this framework can achieve effective decision making, the mapping may disrupt the original geometric structure of the Riemannian features, thereby introducing semantic bias into the decision process (Chen et al., 2024a;b; 2026a; Yu & Xu, 2026).

To address this issue, we explore constructing decision modules on Riemannian manifolds using Multinomial Logistic Regression (MLR). In recent years, several studies have begun to exploit the intrinsic geometry of manifolds to construct classification decisions (Ganea et al., 2018; Chen et al., 2024a;b; 2026a). For instance, in hyperbolic manifolds, Busemann MLR employs Busemann func-

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.

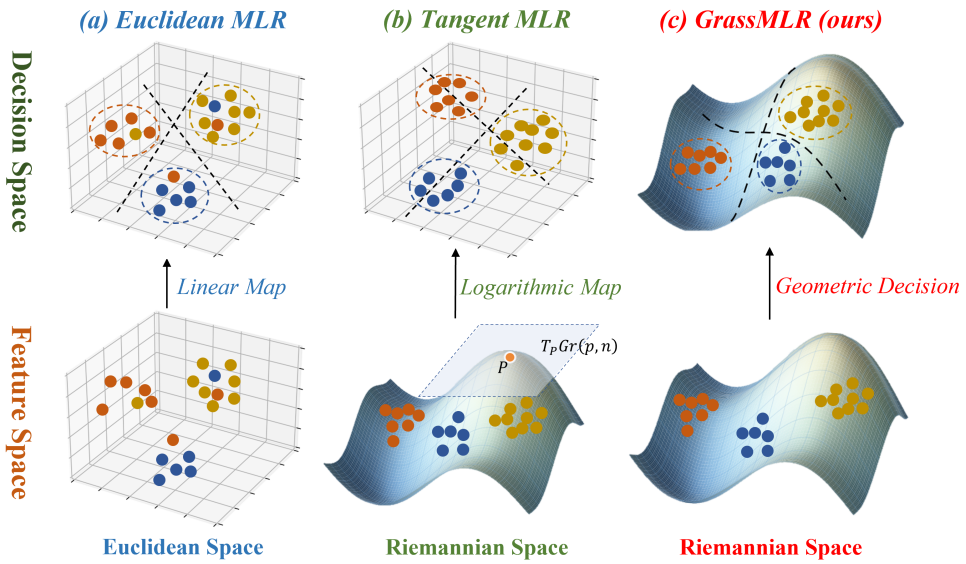

Figure 1: **Paradigm comparisons in feature and decision spaces.** (a) represents features in Euclidean space, failing to capture their intrinsic structure and thereby causing semantic errors and biased decisions. (b) maps Riemannian features to a Euclidean tangent space for classification, disrupting their intrinsic semantics and leading to biased decisions. (c) performs both feature representation and multinomial logistic regression directly on the Grassmann manifold, preserving the intrinsic geometric semantics and achieving geometrically consistent decisions.

tions and their level sets to construct geometrical decision hypersurfaces (Chen et al., 2026a). On symmetric positive definite (SPD) manifolds, SPDMLR and RMLR (Chen et al., 2024a;b) construct decision hypersurfaces on manifolds from the definition of multinomial logistic regression, thereby reducing their reliance on specific manifold geometries. Although these studies provide an important foundation for geometric decision making on manifolds, existing methods typically still rely on tangent-space mappings to the Euclidean during the construction of decision hypersurfaces. This compromise facilitates simple calculation, but inevitably introduces geometric distortion (Chen et al., 2024a;b).

To avoid the negative impact of Euclidean geometry when constructing decision hypersurfaces on the Grassmann manifold, we directly exploit its intrinsic geodesic geometry to construct decision boundaries with explicit discriminative design.

Although the decision boundaries constructed from intrinsic geodesic geometry can avoid the geometric effects introduced by mappings to Euclidean space, subspace representation is also another important factor in obtaining clearer decision boundaries. Existing attempts typically obtain orthogonal subspace representations by applying singular value decomposition to the original data (Ionescu et al., 2015; Huang et al., 2018; Wang et al., 2024; Yu & Xu, 2026). For example, GrNet and GDLNet extract orthogonal bases via SVD and perform subspace representation learning on this basis (Huang et al., 2018; Wang et al., 2024), while GMSF-Net further considers the compatibility among subspaces and adaptively constructs multiple subspaces to enhance the representation capacity (Yu & Xu, 2026). Although these methods enrich subspace representations to some extent, their representation learning processes remain primarily centred on specific orthogonal bases, overlooking the equivalence of different orthogonal bases within a subspace (Edelman et al., 1998; Bendokat et al., 2024). This delivers representation differences unrelated to the subspace itself, thereby limiting the effectiveness of the geometric representations. This naturally raises the question: *How can we jointly construct clearer decision boundaries from both the decision and representation spaces?*

To answer the above question, we first revisit classification decision making from the perspective of the geometric structure of the Grassmann manifold and propose **Grassmann Multinomial Logistic**

Regression (GrassMLR), which aims to directly exploit geodesic distances on the Grassmann manifold to construct classification decision boundaries with explicit geometric guarantees. Specifically, we introduce two learnable Grassmann anchors for each class and compute the squared geodesic distances between an input subspace and the two anchors. We then construct a distance-difference function from these two distances, thereby defining a decision hypersurface with explicit geometric meaning on the Grassmann manifold. Furthermore, to obtain geometric representations better suited to GrassMLR, we propose a **Quotient-Space Representation Learning (QSRL)** strategy. Specifically, we exploit the quotient-space invariance of the Grassmann manifold and introduce a learnable Lie-group transformation to adjust the orthonormal basis representation of each input subspace, while preserving the underlying subspace and its equivalence class, thereby obtaining more effective geometric representations (Edelman et al., 1998; Bendokat et al., 2024).

Our main contributions are listed below:

- We propose GrassMLR, which constructs a distance-difference function to define a decision hypersurface with explicit geometric guarantee on the Grassmann manifold, thereby enabling classification based on its intrinsic geometry.
- We propose a QSRL strategy, which exploits the quotient-space invariance of the Grassmann manifold to adaptively adjust subspace representations within the same equivalence class while preserving the underlying subspaces.
- We conduct extensive experiments on multiple downstream tasks. The results demonstrate that GrassMLR and QSRL work synergistically to improve the performance of Grassmann neural networks and obtain more discriminative geometric representations.

2 Related Work

Grassmann Neural Networks. The Grassmann manifold provides a natural mathematical framework for characterising the intrinsic geometric structure of subspace-valued data by representing data as low-dimensional linear subspaces (Edelman et al., 1998; Bendokat et al., 2024; Nguyen & Yang, 2023; Nguyen et al., 2024). Within this framework, prior studies primarily focused on subspace metrics and discriminative representation learning, exploiting kernel methods and discriminant analysis to capture geometric relationships and discriminative information among subspaces (Hamm & Lee, 2008; Harandi et al., 2013; Huang et al., 2015). With the development of deep learning, GrNet extended subspace learning to multilayer networks through full-rank mapping, re-orthogonalisation, and projection-based operations for feature transformation and aggregation (Huang et al., 2018). GDLNet subsequently introduced Grassmann self-attention, using distances between subspaces to compute attention weights and feature aggregation to model relationships across channels (Wang et al., 2024). GMSF-Net further investigated adaptive multi-subspace construction and fusion, learning complementary information through stackable interaction modules (Yu & Xu, 2026). Although the aforementioned studies have achieved promising results in Grassmann subspace representation learning, representation learning and classification decision-making are typically treated separately, with the learned Riemannian representations still mapped to Euclidean space for classification (Huang et al., 2018; Wang et al., 2024; Yu & Xu, 2026). Our work complements these advances in feature learning by focusing on geodesic distance based classification and adaptive orthonormal basis adjustment.

Multinomial Logistic Regression on Manifolds. Multinomial logistic regression (MLR) in Euclidean space can be interpreted through the signed distance from a sample to a class-specific reference hyperplane, providing a basis for extending classification to Riemannian manifolds (Chen et al., 2024b; 2026a). Following this perspective, existing studies have developed MLR classifiers tailored to the geometric structures of different manifolds. In hyperbolic space, early approaches constructed Poincaré MLR using hyperbolic hyperplanes and point-to-hyperplane distances (Ganea et al., 2018; Shimizu et al., 2021), followed by MLR formulations in the Lorentz model (Bdeir et al., 2024; Nguyen et al., 2025). Busemann MLR instead constructs classification decisions using Busemann functions and their level sets, providing a geometric interpretation in terms of the distance from a point to a horosphere (Chen et al., 2026a). For symmetric positive definite manifolds, related studies have developed SPD MLR using gyrovector structures (Nguyen & Yang, 2023) or metrics induced from Euclidean space (Chen et al., 2024a; 2025; 2026b). Furthermore, RMLR adopts a more general Riemannian perspective and defines class scores through the Riemannian logarithmic map, thereby

reducing reliance on specific manifold structures (Chen et al., 2024b). Although these studies provide an important foundation for geometric classification on manifolds, existing approaches often rely on tangent-space representations to construct classification decisions (Chen et al., 2024a;b). In contrast, GrassMLR defines class reference sets through two-anchor squared geodesic distance and uses gradient normalisation to obtain a local signed-distance interpretation.

3 PRELIMINARIES

This section briefly introduces the Grassmann geometry and multinomial logistic regression. Further details are provided in Appendix D.

Grassmann Manifold. The Grassmann manifold $\mathcal{G}(n,p)$ is the space of all p-dimensional linear subspaces of $\mathbb{R}^n$, with dimension $p(n-p)$. It can be formally defined as

$$\mathcal{G}(n,p) = \{\mathcal{S} \subseteq \mathbb{R}^n \mid \dim(\mathcal{S}) = p\}. \tag{1}$$

Each point on the Grassmann manifold can be represented by an orthonormal basis matrix $U \in \mathbb{R}^{n\times p}$ satisfying $U^\top U = I_p$. All $n \times p$ matrices satisfying the above orthonormality constraint form the Stiefel manifold $\mathrm{St}(n,p)$. Thus, U is an element of the Stiefel manifold $\mathrm{St}(n,p)$. For any orthogonal matrix $Q \in O(p)$, U and UQ provide different orthonormal basis representations but span the same p-dimensional subspace, *i.e.*, $U \sim UQ, Q \in O(p)$. Therefore, the Grassmann manifold can be represented as the quotient of the Stiefel manifold by the orthogonal group $O(p)$:

$$\mathcal{G}(n,p) = \mathrm{St}(n,p)/O(p). \tag{2}$$

The Riemannian metric, logarithmic map, exponential map, and tangent-space norm are denoted by $\langle\cdot,\cdot\rangle_X$, $\mathrm{Log}_X(\cdot)$, $\mathrm{Exp}_X(\cdot)$, and $|\cdot|_X$, respectively. Their definitions and computational forms are provided in Appendix D.1.

Multinomial Logistic Regression. For C classes, Multinomial Logistic Regression (MLR) in $\mathbb{R}^n$ assigns each class a linear logit, which can be geometrically interpreted in terms of the signed distance from the input feature x to a class-specific Euclidean hyperplane:

$$p(y=k \mid x) \propto \exp(\mathrm{sign}(\langle a_k, x-p_k\rangle)\,\|a_k\|\,d(x,\mathcal{H}_{a_k,p_k})), \tag{3}$$

where $a_k \in \mathbb{R}^n \setminus \{0\}$ is the normal vector of the k-th hyperplane, $p_k \in \mathbb{R}^n$ is any point on the hyperplane, and

$$\mathcal{H}_{a_k,p_k} = \{x \in \mathbb{R}^n \mid \langle a_k, x-p_k\rangle = 0\}. \tag{4}$$

Since

$$\langle a_k, x-p_k\rangle = \mathrm{sign}(\langle a_k, x-p_k\rangle)\,\|a_k\|\,d(x,\mathcal{H}_{a_k,p_k}), \tag{5}$$

the exponent is exactly the linear logit $\langle a_k, x-p_k\rangle$. Thus, the magnitude of the logit is determined by the distance to the decision hyperplane, while its sign indicates the side of the hyperplane on which x lies. The class probabilities are then obtained by applying the softmax normalisation across the C class logits.

4 GRASSMANN MULTINOMIAL LOGISTIC REGRESSION(GRASSMLR)

This section provides a detailed introduction to the proposed Grassmann Multinomial Logistic Regression (GrassMLR) method. GrassMLR exploits the intrinsic geodesic geometry of the Grassmann manifold for classification by constructing a distance difference function from the difference in squared geodesic distances to two learnable reference points, and normalising the class scores by their gradient norms, thereby establishing a correspondence with local scaled signed distances.

4.1 EUCLIDEAN HYPERPLANES FROM DISTANCE DIFFERENCES FUNCTION

To establish the geometric interpretation of distance difference functions, consider two distinct reference points $A_k, B_k \in \mathbb{R}^n$ for the k-th class. We define the squared distance difference function as

$$F_k(x) = \|x - A_k\|^2 - \|x - B_k\|^2. \tag{6}$$

Expanding the two squared distances gives

$$\begin{aligned} F_k(x) &= (x - A_k)^\top (x - A_k) - (x - B_k)^\top (x - B_k) \\ &= 2\langle B_k - A_k, x\rangle + \|A_k\|^2 - \|B_k\|^2. \end{aligned} \tag{7}$$

It can be seen that the quadratic terms in x in $F_k(x)$ completely cancel. Therefore, although it is constructed from two squared distances, it ultimately reduces to an affine function of x. In particular, the zero level set

$$\mathcal{H}_k^0 = \{x \in \mathbb{R}^n \mid F_k(x) = 0\} \tag{8}$$

is a Euclidean hyperplane. This hyperplane corresponds precisely to all points that have equal squared distances to A_k and B_k, namely, $\|x - A_k\|^2 = \|x - B_k\|^2$. Therefore, the squared distance difference function of the two reference points not only compares the relative distances between the input and the two reference points, but also naturally induces a hyperplane that partitions the Euclidean space into two regions. Specifically, $F_k(x) > 0$ and $F_k(x) < 0$ indicate that the input is closer to B_k and A_k, respectively.

Furthermore, the gradient of the distance difference function is $\nabla F_k(x) = 2(B_k - A_k)$. Since $A_k \neq B_k$, this gradient norm is a positive constant and is independent of the input x. Therefore, $B_k - A_k$ gives the normal direction of the hyperplane, while $\|\nabla F_k(x)\|$ characterises the rate of change of the distance difference function along this normal direction. To control the score magnitude and select a class reference hyperplane, we introduce a class-specific scale parameter $\alpha_k > 0$ and an offset $b_k \in \mathbb{R}$, and define

$$u_k(x) = \frac{-\alpha_k F_k(x) + b_k}{\|\nabla F_k(x)\|}. \tag{9}$$

The corresponding decision hyperplane is defined as $\mathcal{H}_k = \{x \in \mathbb{R}^n \mid u_k(x) = 0\}$, which is equivalently characterised by $-\alpha_k F_k(x) + b_k = 0$. Expanding $F_k(x)$, we obtain

$$u_k(x) = \left\langle \alpha_k \frac{A_k - B_k}{\|A_k - B_k\|}, x \right\rangle + \frac{b_k - \alpha_k(\|A_k\|^2 - \|B_k\|^2)}{2\|A_k - B_k\|}. \tag{10}$$

Since $\nabla u_k(x) = -\frac{\alpha_k \nabla F_k(x)}{\|\nabla F_k(x)\|}$, we have $\|\nabla u_k(x)\| = \alpha_k$. Therefore, the signed Euclidean distance from the input x to the hyperplane $\mathcal{H}_k$ is given by

$$d_{\text{signed}}(x, \mathcal{H}_k) = \frac{u_k(x)}{\alpha_k}. \tag{11}$$

The signed distance is positive on the side where $u_k(x) > 0$. Thus, $u_k(x)$ is essentially the scaled signed distance to the hyperplane, where α_k controls the distance scale. Using $u_k(x)$ as the class logit, the softmax formulation of multinomial logistic regression takes the standard Euclidean form:

$$p(y = k \mid x) = \frac{\exp(u_k(x))}{\sum_{j=1}^{C} \exp(u_j(x))}. \tag{12}$$

This gives the standard Euclidean logit consisting of a linear term and a bias. Conversely, every nonconstant Euclidean logit admits this parameterisation, as shown in Appendix E.5.

4.2 Grassmann Hypersurfaces From Distance Differences

We extend the squared distance difference construction to the Grassmann manifold. Let $X \in \mathcal{G}(n, p)$ be an input subspace. For each class k, we introduce two distinct learnable anchor subspaces $A_k, B_k \in \mathcal{G}(n, p)$ and define

$$\begin{aligned} F_k(X) = d_{\mathcal{G}}^2(X, A_k) - d_{\mathcal{G}}^2(X, B_k) &= \left\| \lg^{A_k} - \lg^X \right\|^2 - \left\| \lg^{B_k} - \lg^X \right\|^2 \\ &= 2\left\langle \lg^{B_k} - \lg^{A_k}, \lg^X \right\rangle + \left\| \lg^{A_k} \right\|^2 - \left\| \lg^{B_k} \right\|^2. \end{aligned} \tag{13}$$

The function compares the distances from the input to the two anchors. Following the Euclidean MLR formulation, we further use the distance difference function to construct an unnormalised class

score on the Grassmann manifold. Specifically, introducing a learnable scale parameter $\alpha_k > 0$ and an offset $b_k \in \mathbb{R}$, we define the unnormalised class score as $g_k(X) = -\alpha_k F_k(X) + b_k$. Substituting the distance difference function into the above formulation gives

$$g_k(X) = \alpha_k \left[d_{\mathcal{G}}^2(X, B_k) - d_{\mathcal{G}}^2(X, A_k) \right] + b_k. \tag{14}$$

Thus, each class score is directly determined by the difference between the intrinsic squared geodesic distances from the input subspace X to two learnable Grassmann anchors.

When $b_k = 0$, the reference set consists of subspaces equidistant from the two anchors; otherwise, it corresponds to a different level set of F_k. Since the Grassmann manifold is compact, F_k has a bounded range. The reference set is empty if b_k/α_k lies outside this range. For $X_0 \in \mathcal{H}_k$, if g_k is smooth near X_0 and $\operatorname{grad} g_k(X_0) \neq 0$, then $\mathcal{H}_k$ is a smooth hypersurface near X_0. The notation, smooth domain, and interpretation of the class reference set are detailed in Appendix A, while its local hypersurface structure is established in Appendix E.2. Here, $\mathcal{H}_k$ is a reference set for class k, while multiclass predictions depend on the relative values of all final class logits. This score will be normalised in the next subsection to obtain the final GrassMLR logit.

Definition 4.1 (Grassmann reference set)**.** Given distinct anchors $A_k, B_k \in \mathcal{G}(n, p)$, a scale $\alpha_k > 0$, and an offset $b_k \in \mathbb{R}$, the class reference set is

$$\begin{aligned} \mathcal{H}_k &= \{X \in \mathcal{G}(n,p) : g_k(X) = 0\} \\ &= \left\{ X \in \mathcal{G}(n,p) : F_k(X) = \frac{b_k}{\alpha_k} \right\}. \end{aligned} \tag{15}$$

4.3 Grassmann Margins from Gradient-Normalised Logits

Section 4.1 shows that the Euclidean distance-difference function has a constant gradient norm. On the Grassmann manifold, this norm generally varies with the input. We introduce gradient normalisation and show that the normalised logit locally approximates a scaled signed geodesic distance.

We normalise the class score using the Riemannian gradient norm of the distance difference function. Away from the cut loci of the two anchors, the squared distance gradient formula gives

$$\operatorname{grad} F_k(X) = -2 \operatorname{Log}_X(A_k) + 2 \operatorname{Log}_X(B_k), \tag{16}$$

and hence

$$\| \operatorname{grad} F_k(X) \|_X = 2 \left\| \operatorname{Log}_X(B_k) - \operatorname{Log}_X(A_k) \right\|_X. \tag{17}$$

For distinct anchors, this gradient is nonzero on the common smooth domain. Indeed, equality of the two logarithmic-map vectors would imply equality of their images under Exp_X, and therefore $A_k = B_k$. The derivation and the non-vanishing property are detailed in Appendix E.1.

Definition 4.2 (GrassMLR)**.** Where the logarithmic maps to all class anchors are well defined, the GrassMLR logit is

$$\begin{aligned} u_k(X) &= \frac{g_k(X)}{\| \operatorname{grad} F_k(X) \|_X + \varepsilon} \\ &= \frac{-\alpha_k F_k(X) + b_k}{\| \operatorname{grad} F_k(X) \|_X + \varepsilon}, \end{aligned} \tag{18}$$

where $\varepsilon \geq 0$. We use $\varepsilon > 0$ in numerical implementations to prevent small denominators. Conditional class probabilities are

$$p(y = k \mid X) = \frac{\exp(u_k(X))}{\sum_{j=1}^{C} \exp(u_j(X))}. \tag{19}$$

The score depends on both the relative distances to the anchors and the local variation of the distance-difference function.

Proposition 4.3 (signed distance approximation)**.** *Let $X_0 \in \mathcal{H}_k$ lie outside the cut loci of both anchors and satisfy*

$$\| \operatorname{grad} F_k(X_0) \|_{X_0} > 0. \tag{20}$$

Then $\mathcal{H}_k$ is a smooth hypersurface near X_0. In a sufficiently small neighbourhood, let $s_k(X)$ denote the signed geodesic distance to this local hypersurface, with the positive side determined by $g_k(X) > 0$. For $\varepsilon = 0$,

$$u_k(X) = \alpha_k s_k(X) + O\big(s_k(X)^2\big). \tag{21}$$

Thus, the unstabilised logit approximates the local signed geodesic distance scaled by α_k to first order. The general normalised-function result is established in Appendix E.3, and the proof specialised to GrassMLR is provided in Appendix E.4.

For $\varepsilon > 0$, in the same neighbourhood,

$$u_k(X) = \frac{\|\operatorname{grad} F_k(X)\|_X}{\|\operatorname{grad} F_k(X)\|_X + \varepsilon} \left[\alpha_k s_k(X) + O\big(s_k(X)^2\big)\right]. \tag{22}$$

The additional factor is close to one when ε is much smaller than the local gradient norm. The corresponding local scaling and error bound are derived in Appendix E.4.

5 Quotient-Space Representation Learning(QSRL)

In addition to constructing classification decisions directly from the intrinsic geometry of the Grassmann manifold, we further exploit its quotient-space structure to adaptively adjust its representation, thereby obtaining geometric representations better suited to GrassMLR. The Grassmann manifold represents a p-dimensional subspace rather than a unique orthonormal basis. Specifically, it can be expressed as

$$\mathcal{G}(n,p) = \mathrm{St}(n,p)/O(p), \tag{23}$$

where $\mathrm{St}(n,p)$ denotes the Stiefel manifold and $O(p)$ denotes the orthogonal group. Based on this quotient space invariance, we introduce a learnable Lie-group transformation $R(U) \in \mathrm{SO}(p)$ to adjust the concrete orthonormal basis representation without changing the underlying Grassmann subspace. For an input basis U, the transformed representation is given by

$$\widetilde{U} = UR(U), \qquad R(U) \in \mathrm{SO}(p). \tag{24}$$

Here, $\mathrm{SO}(p)$ is the special orthogonal group. Since $R(U)$ is orthogonal, we have $R(U)^\top R(U) = I_p$. Consequently, $\widetilde{U}\widetilde{U}^\top = UR(U)R(U)^\top U^\top = UU^\top$, showing that the transformation preserves the underlying Grassmann subspace. The transformation $R(U)$ is generated adaptively for each input, allowing the basis representation to be adjusted while preserving the underlying Grassmann geometry. The resulting representation $\widetilde{U}$ is then fed into the GrassMLR classifier for intrinsic geometric classification. The relevant operations and properties of QSRL within GraSS are provided in Appendix E.6 and Appendix E.9.

In this way, QSRL and GrassMLR address complementary aspects of Grassmannian learning. QSRL optimises the representative of each subspace within its quotient equivalence class, while GrassMLR constructs class-specific decision boundaries directly from the intrinsic geometry of the resulting subspaces. Their combination therefore enables both representation adaptation and geometry-aware classification without changing the underlying subspace structure.

Proposition 5.1 (Quotient Space Invariance)**.** *For any $U \in \mathrm{St}(n,p)$ and $Q \in O(p)$, the transformed basis $\widetilde{U} = UQ$ belongs to the same quotient-space equivalence class as U, i.e.,*

$$[\widetilde{U}] = [U], \qquad \widetilde{U}\widetilde{U}^\top = UU^\top. \tag{25}$$

Therefore, QSRL changes only the concrete basis representation without altering the corresponding Grassmann point.

6 Experiments

We evaluate GrassMLR on human action recognition and hyperspectral image classification using FPHA, HDM05, and Salinas. In addition to classification performance, we report the model size and floating-point operations (FLOPs) to further examine the computational cost introduced by GrassMLR.

Table 1: Comparison with existing Riemannian methods on the HDM05 and FPHA datasets.

Method	HDM05			FPHA		
	Acc. (%)	Size (MB)	FLOPs (M)	Acc. (%)	Size (MB)	FLOPs (M)
SPDNet (Huang & Gool, 2017)	$60.45\% \pm 1.12$	11.71	2050.49	$85.65\% \pm 1.02$	13.60	1595.50
MATT (Pan et al., 2022)	$\mathbf{62.25\% \pm 1.68}$	3.99	442.45	$\underline{86.70\% \pm 0.68}$	$\underline{1.83}$	142.07
GDLNet (Wang et al., 2024)	$60.08\% \pm 1.78$	$\underline{3.95}$	89.11	$86.60\% \pm 0.69$	$\underline{1.83}$	33.69
GrNet (Huang et al., 2018)	$59.23\% \pm 1.78$	6.88	**81.96**	$78.79\% \pm 1.82$	6.73	**23.60**
GrassMLR (Ours)	$\underline{61.49\% \pm 0.79}$	**2.75**	$\underline{88.87}$	$\mathbf{86.97\% \pm 0.46}$	**0.84**	$\underline{26.56}$

6.1 Experimental Setup

Datasets. We use FPHA (Garcia-Hernando et al., 2018) and HDM05 (Müller et al., 2007) for human action recognition, and Salinas (Green et al., 1998; Grupo de Inteligencia Computacional) for hyperspectral image classification. These datasets provide temporal action representations and local spectral representations on Grassmann spaces of different dimensions. The dataset preparation and partitions are detailed in Appendix F.1.

Comparison Protocol. To comprehensively evaluate the effectiveness of GrassMLR, we first compare it with existing Riemannian neural networks, including SPDNet, MATT, GrNet, GDLNet, and GMSF-Net. For each paired comparison, we keep the data split, preprocessing, feature-extraction network, and optimisation strategy unchanged, and only replace the final classification module to compare GrassMLR with the Euclidean classifier under the same feature representation. Furthermore, we evaluate the effect of QSRL on the performance of GrassMLR. Details of the network architectures, training settings, and specific configurations of the compared methods are provided in Appendix F.2.

Evaluation. We report classification accuracy on FPHA and HDM05, with each experiment repeated 10 times using different random seeds; the mean and standard deviation are reported. For Salinas, we report overall accuracy (OA), average accuracy (AA), and the Kappa coefficient. We also report model size and FLOPs. The best and second-best results are highlighted in bold and underlined, respectively. Further evaluation details are given in Appendix F.2.

6.2 Experimental Analysis

Human Action Recognition. As shown in Table 1, GrassMLR achieves strong classification performance on both the HDM05 and FPHA datasets, obtaining accuracies of $61.49\% \pm 0.79$ and $86.97\% \pm 0.46$, respectively. Compared with existing Grassmann methods, namely GrNet and GDLNet, GrassMLR achieves higher classification accuracy on both datasets. In terms of model size, GrassMLR has the smallest model size on both datasets, requiring only 2.75 MB and 0.84 MB, respectively. Compared with GrNet, GrassMLR reduces the model size by approximately 60.0% and 87.5% on HDM05 and FPHA, respectively, while improving classification accuracy by 2.26 and 8.18 percentage points. In terms of computational cost, GrassMLR requires 88.87M and 26.56M FLOPs on HDM05 and FPHA, respectively, corresponding to increases of 6.91M and 2.96M FLOPs over GrNet. Overall, GrassMLR achieves the highest accuracy on FPHA and the second-highest accuracy on HDM05, while maintaining the smallest model size on both datasets.

Hyperspectral Image Classification. We further evaluate GrassMLR on the Salinas dataset to examine its applicability to Grassmann representations constructed from local spectral features. Using the checkpoint selected by validation OA, GrassMLR achieves a test OA of 97.54%, an AA of 98.18%, and a Kappa coefficient of 0.9726. As shown in Table 2, GrassMLR achieves the highest OA, AA, and Kappa among all compared methods, improving OA and AA by 0.72 and 1.07 percentage points, respectively, and Kappa by 0.0044 over the second-best result for each metric. Moreover, GrassMLR has the smallest model size at only 0.0827 MB, which is approximately 20.1% smaller than GrNet, the second-smallest model. In terms of computational cost, GrassMLR requires 1.40 MFLOPs, the second-lowest computational cost among the compared methods. These results demonstrate that GrassMLR achieves the best classification performance while using the smallest model size, with a relatively low computational cost.

Table 2: Performance comparison of Riemannian methods on the Salinas dataset.

Method	OA	AA	Kappa	Size (MB)	FLOPs (M)
SPDNet (Huang & Gool, 2017)	92.72%	94.86%	0.9286	0.1237	2.88
MATT (Pan et al., 2022)	90.03%	91.78%	0.8826	0.8563	3.58
GrNet (Huang et al., 2018)	96.51%	<u>97.11%</u>	0.9612	<u>0.1035</u>	**1.08**
GDLNet (Wang et al., 2024)	95.46%	94.33%	0.9563	0.1436	1.58
GMSF-Net (Yu & Xu, 2026)	<u>96.82%</u>	96.73%	<u>0.9682</u>	1.5868	2.63
GrassMLR (Ours)	**97.54%**	**98.18%**	**0.9726**	**0.0827**	<u>1.40</u>

Table 3: Ablation study of GrassMLR and QSRL on the HDM05 and FPHA datasets using a fixed random seed.

Method	GrassMLR	QSRL	HDM05	FPHA
GrNet	×	×	59.37%	79.88%
GrNet + GrassMLR	✓	×	<u>60.71%</u>	<u>86.78%</u>
GrNet + QSRL	×	✓	60.32%	85.39%
GrNet + GrassMLR + QSRL	✓	✓	**61.49%**	**86.97%**

6.3 Ablation Studies

To further analyse the individual contributions of GrassMLR and QSRL, we conduct an ablation study on the HDM05 and FPHA datasets, with the results presented in Table 3. All ablation configurations are evaluated using a fixed random seed. The GrNet baseline is evaluated under this fixed-seed setting, accounting for its different results from those in Table 1.

First, we use the original GrNet as the baseline, which achieves accuracies of 59.37% and 79.88% on HDM05 and FPHA, respectively. With QSRL disabled, introducing GrassMLR increases the accuracies to 60.71% and 86.78%, corresponding to improvements of 1.34 and 6.90 percentage points over the baseline, respectively. This demonstrates that classification based on the intrinsic geodesic geometry of the Grassmann manifold can effectively improve classification performance. On the other hand, when retaining the original classifier and introducing only QSRL, the accuracies reach 60.32% and 85.39% on HDM05 and FPHA, respectively, yielding improvements of 0.95 and 5.51 percentage points over the baseline. This indicates that quotient-space representation learning can further improve classification performance by adjusting the Grassmann representations.

Furthermore, when GrassMLR and QSRL are introduced jointly, the model improves upon the original GrNet by 2.12 and 7.09 percentage points on HDM05 and FPHA, respectively. The joint configuration also consistently outperforms the configurations using GrassMLR or QSRL alone, indicating that the two strategies can jointly improve the classification performance of Grassmann neural networks from the perspectives of classification decision and representation learning.

7 Conclusion

This paper proposes Grassmann Multinomial Logistic Regression (GrassMLR) and a Quotient Space Representation Learning (QSRL) strategy, providing a geometry aware unified framework for classification decision making and representation learning on Grassmann manifolds. Specifically, GrassMLR constructs decision hypersurfaces based on the differences between squared geodesic distances from an input subspace to learnable Grassmann anchors, thereby ensuring that classification decisions remain consistent with the underlying manifold geometry. Furthermore, QSRL exploits the quotient space invariance of the Grassmann manifold to adaptively transform orthonormal basis representations while preserving the underlying subspaces and their equivalence classes, thereby obtaining more effective geometric representations for decision making. Overall, this work extends Grassmannian learning from geometry aware representation learning to the joint modelling of geometry aware decision making and quotient-space representation learning.

Ethics Statement

All experiments in this work are conducted on publicly available benchmark datasets (HDM05, FPHA, and Salinas), which contain no personally identifiable or sensitive information. Therefore, this work does not raise ethical issues.

Reproducibility Statement

This work is designed to be fully reproducible. We provide comprehensive descriptions of all datasets, experimental protocols, and evaluation metrics in the experiments section, while additional implementation details and training parameters are deferred to the supplementary material. The source code and configuration files will be released upon acceptance to support reproduction and follow-up studies.

References

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ekz1hN5QNh.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7.

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50 (1):6, 2024. doi: 10.1007/s10444-023-10090-8. URL https://doi.org/10.1007/s10444-023-10090-8.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiao-Jun Wu, and Nicu Sebe. Riemannian multinomial logistics regression for spd neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17086–17096, June 2024a.

Ziheng Chen, Yue Song, Rui Wang, Xiaojun Wu, and Nicu Sebe. RMLR: Extending multinomial logistic regression into general geometries. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=lBp2cda7sp.

Ziheng Chen, Yue Song, Xiaojun Wu, Gaowen Liu, and Nicu Sebe. Understanding matrix function normalizations in covariance pooling through the lens of riemannian geometry. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=q1t0Lmvhty.

Ziheng Chen, Bernhard Schölkopf, and Nicu Sebe. Hyperbolic busemann neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42029–42038, June 2026a.

Ziheng Chen, Yue Song, Xiaojun Wu, and Nicu Sebe. Fast and stable riemannian metrics on SPD manifolds via cholesky product geometry. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=5S8ruWKe8l.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/dbab2adc8f9d078009ee3fa810bea142-Paper.pdf.

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–419, 2018.

Robert O. Green, Michael L. Eastwood, Charles M. Sarture, Thomas G. Chrien, Mikael Aronsson, Bruce J. Chippendale, Jessica A. Faust, Betina E. Pavri, Christopher J. Chovit, Manuel Solis, Martin R. Olah, and Orlesa Williams. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 65(3):227–248, 1998. doi: 10.1016/S0034-4257(98)00064-9.

Grupo de Inteligencia Computacional. Hyperspectral remote sensing scenes: Salinas. https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes. Accessed: 2026-09-18.

Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 376–383, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390204. URL https://doi.org/10.1145/1390156.1390204.

Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *2013 IEEE International Conference on Computer Vision*, pp. 3120–3127, 2013. doi: 10.1109/ICCV.2013.387.

Zhiwu Huang and Luc Van Gool. A riemannian network for SPD matrix learning. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 2036–2042. AAAI Press, 2017. doi: 10.1609/AAAI.V31I1.10866. URL https://doi.org/10.1609/aaai.v31i1.10866.

Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on Grassmann manifolds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, volume 32, pp. 1137–1145, 2018. doi: 10.1609/aaai.v32i1.11725. URL https://ojs.aaai.org/index.php/AAAI/article/view/11725.

Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, 2007.

Xuan Son Nguyen and Shuo Yang. Building neural networks on matrix manifolds: A gyrovector space approach. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26031–26062. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nguyen23f.html.

Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Matrix manifold neural networks++. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=30aSE3FB3L.

Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Neural networks on symmetric spaces of noncompact type. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bwOndfohRK.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. Matt: A manifold attention network for eeg decoding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31116–31129. Curran Associates, Inc., 2022. doi: 10.52202/068431-2256. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c981fd12b1d5703f19bd8289da9fc996-Paper-Conference.pdf.

Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ec85b0tUwbA.

Rui Wang, Chen Hu, Ziheng Chen, Xiao-Jun Wu, and Xiaoning Song. A grassmannian manifold self-attention network for signal classification. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5099–5107. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/564. URL https://doi.org/10.24963/ijcai.2024/564. Main Track.

Xuan Yu and Tianyang Xu. Learning topology-driven multi-subspace fusion for grassmannian deep networks. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 27987–27995. AAAI Press, 2026. doi: 10.1609/AAAI.V40I33.40023. URL https://doi.org/10.1609/aaai.v40i33.40023.

APPENDIX

A Default Notation 14
B Limitations 14
C AI Use Statement 14
D Preliminaries 14
D.1 Grassmannian 14
E Proof 15
E.1 Proof of Proposition E.1 15
E.2 Proof of Proposition E.2 16
E.3 Proof of Proposition E.3 17
E.4 Proof of Proposition 4.3 18
E.5 Proof of Proposition E.4 19
E.6 Proof of Proposition E.5 20
E.7 Proof of Proposition E.6 21
E.8 Proofs of Propositions E.7 and E.8 21
E.9 Proof of Proposition 5.1 22
E.10 Proof of Proposition E.11 23
E.11 Proof of Proposition E.12 24
F Experiments 24
F.1 Datasets 24
F.2 Implementation Details 25

A Default Notation

Let $\mathcal{M} = \mathcal{G}(n,p)$ be equipped with the canonical Grassmann metric, where $1 \leq p < n$. We write $d_{\mathcal{G}}$ for its geodesic distance, $\langle \cdot, \cdot \rangle_X$ for the Riemannian inner product at X, and $\| \cdot \|_X$ for the associated norm. Throughout this appendix, grad denotes the Riemannian gradient, while ∇ denotes the Euclidean gradient.

For distinct anchors $A, B \in \mathcal{M}$, define

$$F_{A,B}(X) = d^2_{\mathcal{G}}(X, A) - d^2_{\mathcal{G}}(X, B). \tag{26}$$

Given a scale $\alpha > 0$ and an offset $b \in \mathbb{R}$, define

$$g_{A,B,\alpha,b}(X) = -\alpha F_{A,B}(X) + b \tag{27}$$

and the class reference set

$$\mathcal{H}_{A,B,\alpha,b} = \{X \in \mathcal{M} : g_{A,B,\alpha,b}(X) = 0\}. \tag{28}$$

When the parameters are clear, we abbreviate these objects as F, g, and $\mathcal{H}$.

Let $\mathrm{Cut}(Y)$ denote the cut locus of Y. The differential calculations involving both anchors are performed on

$$\mathcal{D} = \mathcal{M} \setminus (\mathrm{Cut}(A) \cup \mathrm{Cut}(B)). \tag{29}$$

On this domain, both squared-distance functions are smooth and the corresponding logarithmic maps are well defined (Boumal, 2023; Bendokat et al., 2024).

The set $\mathcal{H}$ is a class reference set rather than, in general, a pairwise multiclass decision boundary. The pairwise decision boundary between classes k and j is determined by the equality of their logits, $u_k(X) = u_j(X)$. The local hypersurface structure of $\mathcal{H}$ is established in Appendix E.2. The local results apply at points of $\mathcal{H} \cap \mathcal{D}$ and do not assume that this intersection is nonempty for every parameter choice.

B Limitations

This work has several limitations. First, the signed-distance interpretation of GrassMLR is a first-order approximation near a regular reference hypersurface and requires the input to lie outside the cut loci of the anchors. The numerical stability term used in practice introduces an additional local scaling, so the logit should not be interpreted as the exact geodesic distance to the reference hypersurface globally. Second, the experiments cover two action-recognition datasets and one hyperspectral dataset; GrassMLR does not achieve the highest action-recognition accuracy in every comparison. Future work should examine the method on a wider range of tasks under unified comparison protocols.

C AI Use Statement

Generative AI tools were used to assist with the organisation and language of the manuscript, primarily to improve the clarity, fluency, and readability of the text. Beyond these language-related tasks, the research ideas, theoretical derivations, methodological design, experimental validation, and associated figures and tables were developed and completed independently by the authors. The authors take full responsibility for the content and final outcomes of this work.

D Preliminaries

D.1 Grassmannian

The Grassmann manifold can be described using orthonormal basis matrices while identifying all bases that span the same subspace (Edelman et al., 1998; Bendokat et al., 2024). Let

$$\mathrm{St}(n,p) = \left\{U \in \mathbb{R}^{n \times p} : U^\top U = I_p\right\} \tag{30}$$

be the Stiefel manifold, and let

$$O(p) = \left\{Q \in \mathbb{R}^{p\times p} : Q^\top Q = I_p\right\} \tag{31}$$

be the orthogonal group. The group $O(p)$ acts on $\mathrm{St}(n,p)$ from the right through $U \mapsto UQ$. We define

$$U \sim V \quad \Longleftrightarrow \quad V = UQ \text{ for some } Q \in O(p). \tag{32}$$

The equivalence class, or orbit, of U is

$$[U] = \{UQ : Q \in O(p)\}. \tag{33}$$

Every matrix in $[U]$ is an orthonormal basis of the same p-dimensional subspace of $\mathbb{R}^n$. The Grassmann manifold therefore admits the quotient representation

$$\mathcal{G}(n,p) \cong \mathrm{St}(n,p)/O(p). \tag{34}$$

Let $\pi : \mathrm{St}(n,p) \to \mathcal{G}(n,p)$ be the quotient map, with $\pi(U) = [U]$. For a Grassmann point $X = [U]$, the corresponding fibre is

$$\pi^{-1}(X) = [U] = \{UQ : Q \in O(p)\}. \tag{35}$$

Thus, U is a particular matrix representative, whereas $X = [U]$ is the geometric point represented by the whole equivalence class. The projection representation makes this invariance explicit: for every $Q \in O(p)$,

$$P_{UQ} = (UQ)(UQ)^\top = UU^\top = P_U. \tag{36}$$

Conversely, two orthonormal matrices with the same projection matrix span the same subspace and hence belong to the same equivalence class.

We next review the geometric quantities used by GrassMLR. Let $X = [U]$ and $Y = [V]$, where $U, V \in \mathrm{St}(n,p)$. Their principal angles $\theta_1, \ldots, \theta_p \in [0, \pi/2]$ are determined by

$$\sigma_i(U^\top V) = \cos\theta_i, \qquad i = 1, \ldots, p, \tag{37}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value. Under the canonical Grassmann metric, the geodesic distance is

$$d_{\mathcal{G}}(X,Y) = \left(\sum_{i=1}^{p} \theta_i^2\right)^{1/2}. \tag{38}$$

The singular values of $U^\top V$ are unchanged under $U \mapsto UQ_1$ and $V \mapsto VQ_2$, where $Q_1, Q_2 \in O(p)$. Therefore, the principal angles and geodesic distance depend only on the quotient points X and Y, rather than on their selected representatives.

We denote the Riemannian norm at X by $\|\cdot\|_X$. When Y lies outside the cut locus of X, $\mathrm{Log}_X(Y)$ is the initial velocity of the unique minimising geodesic from X to Y, parameterised on $[0,1]$, and satisfies

$$\|\mathrm{Log}_X(Y)\|_X = d_{\mathcal{G}}(X,Y). \tag{39}$$

For a fixed Y and differentiation with respect to X, the squared geodesic distance is smooth away from the cut locus and satisfies (Boumal, 2023; Bendokat et al., 2024)

$$\mathrm{grad}_X\, d^2_{\mathcal{G}}(X,Y) = -2\,\mathrm{Log}_X(Y). \tag{40}$$

A derivation based on the first variation formula is provided in Appendix E.1. Throughout the paper, grad denotes a Riemannian gradient, whereas ∇ denotes a Euclidean gradient.

E Proof

E.1 Proof of Proposition E.1

Proposition E.1 (Gradient of the squared distance)**.** *For $Y \in \mathcal{M}$, consider the squared-distance function with respect to the input variable X:*

$$\phi_Y(X) = \frac{1}{2} d^2_{\mathcal{G}}(X,Y). \tag{41}$$

When taking the gradient with respect to X, the point Y is held constant. For $X \notin \mathrm{Cut}(Y)$,

$$\mathrm{grad}_X\, \phi_Y(X) = -\,\mathrm{Log}_X(Y). \tag{42}$$

Hence,

$$\mathrm{grad}_X\, d^2_{\mathcal{G}}(X,Y) = -2\,\mathrm{Log}_X(Y). \tag{43}$$

In GrassMLR, Y corresponds to a learnable anchor; it is held constant when differentiating with respect to the input X.

Proof. Let $v \in T_X\mathcal{M}$ be arbitrary and choose a smooth curve $\gamma(t)$ satisfying

$$\gamma(0) = X, \qquad \dot{\gamma}(0) = v. \tag{44}$$

Let $\sigma : [0,1] \to \mathcal{M}$ be the minimising geodesic from X to Y, parameterised with constant speed. With this parameterisation,

$$\sigma(0) = X, \qquad \sigma(1) = Y, \qquad \dot{\sigma}(0) = \operatorname{Log}_X(Y). \tag{45}$$

The speed is $d_{\mathcal{G}}(X, Y)$.

The first variation formula for geodesic energy gives (Boumal, 2023)

$$\begin{aligned} \frac{d}{dt}\Big|_{t=0} \frac{1}{2} d^2_{\mathcal{G}}(\gamma(t), Y) &= -\langle v, \dot{\sigma}(0) \rangle_X \\ &= -\langle v, \operatorname{Log}_X(Y) \rangle_X . \end{aligned} \tag{46}$$

By the definition of the Riemannian gradient,

$$\frac{d}{dt}\Big|_{t=0} \phi_Y(\gamma(t)) = \langle \operatorname{grad}_X \phi_Y(X), v \rangle_X . \tag{47}$$

Comparing these identities for every $v \in T_X\mathcal{M}$ yields

$$\operatorname{grad}_X \phi_Y(X) = -\operatorname{Log}_X(Y). \tag{48}$$

Multiplying by two proves the second identity.

Applying Proposition E.1 to the two anchors and using linearity of the gradient gives

$$\begin{aligned} \operatorname{grad} F(X) &= -2 \operatorname{Log}_X(A) + 2 \operatorname{Log}_X(B) \\ &= 2 \left(\operatorname{Log}_X(B) - \operatorname{Log}_X(A) \right). \end{aligned} \tag{49}$$

Therefore,

$$\| \operatorname{grad} F(X) \|_X = 2 \left\| \operatorname{Log}_X(B) - \operatorname{Log}_X(A) \right\|_X . \tag{50}$$

In an orthonormal basis representation, the Riemannian norm can be evaluated as the Frobenius norm of the corresponding horizontal tangent matrix (Bendokat et al., 2024).

For distinct anchors, this gradient is nonzero on $\mathcal{D}$. Indeed, if $\operatorname{grad} F(X) = 0$, then

$$\operatorname{Log}_X(A) = \operatorname{Log}_X(B). \tag{51}$$

Applying Exp_X to both sides gives $A = B$, which contradicts the assumption that the anchors are distinct.

The gradient norm generally depends on X and is not constrained to equal one. Its nonvanishing on $\mathcal{D}$ does not by itself provide a uniform positive lower bound over all inputs and parameter choices. In particular, the value $|F(X) - \tau|$ cannot generally be identified with the exact distance to the level set $\{X : F(X) = \tau\}$.

E.2 Proof of Proposition E.2

Proposition E.2 (Local reference hypersurface)**.** *Let $X_0 \in \mathcal{H} \cap \mathcal{D}$ satisfy*

$$\operatorname{grad} F(X_0) \neq 0. \tag{52}$$

Then there exists a neighbourhood $\mathcal{U}$ of X_0 such that $\mathcal{H} \cap \mathcal{U}$ is a smooth hypersurface of dimension

$$\dim(\mathcal{H} \cap \mathcal{U}) = p(n-p) - 1. \tag{53}$$

Proof. Since X_0 lies outside the cut loci of both anchors, the squared-distance functions are smooth near X_0. Thus, F and $g = -\alpha F + b$ are smooth in a neighbourhood of X_0.

The differential of g satisfies

$$dg_{X_0} = -\alpha\, dF_{X_0}. \tag{54}$$

Equivalently,

$$\operatorname{grad} g(X_0) = -\alpha \operatorname{grad} F(X_0). \tag{55}$$

Since $\alpha > 0$ and $\operatorname{grad} F(X_0) \neq 0$, the differential dg_{X_0} is nonzero. Because its codomain is one-dimensional, it is surjective.

The implicit function theorem therefore implies that the zero level set of g is a smooth hypersurface near X_0 (Boumal, 2023). Its codimension is one. As $\dim \mathcal{G}(n,p) = p(n-p)$, its dimension is $p(n-p) - 1$.

This is a local statement. It does not assert that the entire reference set is smooth across the cut loci. The argument uses smoothness of the squared distance rather than smoothness of the unsquared distance at an anchor.

E.3 Proof of Proposition E.3

Proposition E.3 (Normalised function value and signed distance)**.** *Let f be a twice continuously differentiable real-valued function near $X_0 \in \mathcal{M}$, with*

$$f(X_0) = 0, \qquad \operatorname{grad} f(X_0) \neq 0. \tag{56}$$

Let $\mathcal{H}_f = \{X : f(X) = 0\}$ denote its local zero level set. In a sufficiently small neighbourhood of X_0, let $s(X)$ be the signed geodesic distance to this local hypersurface, with its positive side determined by $f(X) > 0$. Then

$$\frac{f(X)}{\|\operatorname{grad} f(X)\|_X} = s(X) + O\big(s(X)^2\big). \tag{57}$$

Proof. By the implicit function theorem, $\mathcal{H}_f$ is a C^2 hypersurface near X_0. After restricting to a sufficiently small normal neighbourhood, each point X has a unique local nearest point $Y \in \mathcal{H}_f$ (Boumal, 2023). Define the positive unit normal at Y by

$$\nu(Y) = \frac{\operatorname{grad} f(Y)}{\|\operatorname{grad} f(Y)\|_Y}. \tag{58}$$

The normal geodesic

$$\gamma_Y(t) = \operatorname{Exp}_Y(t\nu(Y)) \tag{59}$$

satisfies $X = \gamma_Y(s)$, where $s = s(X)$. Positive values of s correspond locally to $f(X) > 0$.

Set

$$a(Y) = \|\operatorname{grad} f(Y)\|_Y > 0. \tag{60}$$

Since $f(Y) = 0$, the first derivative along the normal geodesic is

$$\begin{aligned} \frac{d}{dt}\bigg|_{t=0} f(\gamma_Y(t)) &= \langle \operatorname{grad} f(Y), \nu(Y)\rangle_Y \\ &= a(Y). \end{aligned} \tag{61}$$

Let $\operatorname{Hess} f$ denote the Riemannian Hessian and write

$$h(Y) = \operatorname{Hess} f(Y)[\nu(Y), \nu(Y)]. \tag{62}$$

Because γ_Y is a geodesic, its covariant acceleration is zero, and hence

$$\frac{d^2}{dt^2}\bigg|_{t=0} f(\gamma_Y(t)) = h(Y). \tag{63}$$

Since $f \in C^2$, the second-order Taylor expansion gives

$$f(X) = a(Y)s + \frac{1}{2}h(Y)s^2 + o(s^2). \tag{64}$$

Next, define

$$q(X) = \| \operatorname{grad} f(X)\|_X. \tag{65}$$

Where the gradient is nonzero, q is continuously differentiable. Metric compatibility gives

$$\begin{aligned} dq_Y[\nu(Y)] &= \frac{\operatorname{Hess} f(Y)[\nu(Y), \operatorname{grad} f(Y)]}{a(Y)} \\ &= h(Y). \end{aligned} \tag{66}$$

Therefore,

$$q(X) = a(Y) + h(Y)s + o(|s|). \tag{67}$$

Since $\operatorname{grad} f(X_0) \neq 0$, continuity allows us to choose the neighbourhood so that

$$q(X) \geq m_f > 0. \tag{68}$$

Restricting the neighbourhood further if necessary also bounds the local second derivatives.

Dividing Equations equation 64 and equation 67 yields

$$\begin{aligned} \frac{f(X)}{q(X)} &= s\frac{1 + \dfrac{h(Y)}{2a(Y)}s + o(|s|)}{1 + \dfrac{h(Y)}{a(Y)}s + o(|s|)} \\ &= s - \frac{h(Y)}{2a(Y)}s^2 + o(s^2) \\ &= s + O(s^2). \end{aligned} \tag{69}$$

The positive local lower bound and bounded second derivatives ensure that the final estimate holds uniformly after restricting the neighbourhood. This proves the result.

E.4 Proof of Proposition 4.3

This subsection proves Proposition 4.3 in the main text and describes the effect of the numerical stability term.

For one class, let F, $g = -\alpha F + b$, and $\mathcal{H} = \{X : g(X) = 0\}$ be as defined above. To distinguish the theoretical and stabilised expressions, write

$$u_\varepsilon(X) = \frac{g(X)}{\| \operatorname{grad} F(X)\|_X + \varepsilon}, \qquad \varepsilon \geq 0. \tag{70}$$

The notation u_0 refers to the expression without the stability term.

Proof of Proposition 4.3. Let $X_0 \in \mathcal{H} \cap \mathcal{D}$ satisfy $\operatorname{grad} F(X_0) \neq 0$. By Proposition E.2, $\mathcal{H}$ is a smooth hypersurface near X_0. Let $s(X)$ be the local signed geodesic distance, positive on the side where $g(X) > 0$.

Applying Proposition E.3 to g gives

$$\frac{g(X)}{\| \operatorname{grad} g(X)\|_X} = s(X) + O\big(s(X)^2\big). \tag{71}$$

Since α and b do not depend on the input X,

$$\operatorname{grad} g(X) = -\alpha \operatorname{grad} F(X). \tag{72}$$

Taking norms and using $\alpha > 0$ yields

$$\| \operatorname{grad} g(X)\|_X = \alpha \| \operatorname{grad} F(X)\|_X. \tag{73}$$

Consequently,

$$\begin{aligned} u_0(X) &= \frac{g(X)}{\|\operatorname{grad} F(X)\|_X} \\ &= \alpha \frac{g(X)}{\|\operatorname{grad} g(X)\|_X} \\ &= \alpha s(X) + O\big(s(X)^2\big). \end{aligned} \tag{74}$$

Restoring the class index gives the statement in the main text.

The positive sign in Equation equation 74 follows from choosing the normal direction along $\operatorname{grad} g$. Reversing the signed-distance convention reverses the sign of its leading term.

Effect of the stability term. Let

$$q_F(X) = \|\operatorname{grad} F(X)\|_X. \tag{75}$$

In the same local neighbourhood,

$$\begin{aligned} u_\varepsilon(X) &= \frac{q_F(X)}{q_F(X) + \varepsilon} u_0(X) \\ &= \frac{q_F(X)}{q_F(X) + \varepsilon} \left[\alpha s(X) + O\big(s(X)^2\big)\right]. \end{aligned} \tag{76}$$

Thus, a fixed positive ε changes the local scaling and cannot generally be omitted from the first-order interpretation.

More precisely, choose a neighbourhood where $q_F(X) \geq m_F > 0$. Since $u_0(X) = O(|s(X)|)$, there exists a local constant C_0 such that

$$\begin{aligned} |u_\varepsilon(X) - u_0(X)| &= \frac{\varepsilon}{q_F(X) + \varepsilon} |u_0(X)| \\ &\leq C_0 \frac{\varepsilon}{m_F} |s(X)|. \end{aligned} \tag{77}$$

Hence,

$$u_\varepsilon(X) = \alpha s(X) + O\big(s(X)^2\big) + O\left(\frac{\varepsilon}{m_F} |s(X)|\right). \tag{78}$$

When ε is small relative to the local gradient norm, the scaling factor in Equation equation 76 is close to one. The result remains a local approximation rather than a global exact distance formula.

E.5 Proof of Proposition E.4

Proposition E.4 (Euclidean correspondence)**.** *Let $a_1, a_2 \in \mathbb{R}^n$ be distinct, $\alpha > 0$, and $b \in \mathbb{R}$. Define*

$$F(x) = \|x - a_1\|^2 - \|x - a_2\|^2. \tag{79}$$

For $\varepsilon = 0$, the normalised score

$$u_0(x) = \frac{-\alpha F(x) + b}{\|\nabla F(x)\|} \tag{80}$$

has the Euclidean MLR form

$$u_0(x) = \langle w, x \rangle + c, \tag{81}$$

with

$$w = \alpha \frac{a_1 - a_2}{\|a_1 - a_2\|}, \qquad c = \frac{b - \alpha(\|a_1\|^2 - \|a_2\|^2)}{2\|a_1 - a_2\|}. \tag{82}$$

Proof. Expanding the squared distances gives

$$F(x) = 2\langle a_2 - a_1, x \rangle + \|a_1\|^2 - \|a_2\|^2. \tag{83}$$

Therefore,

$$\nabla F(x) = 2(a_2 - a_1), \qquad \|\nabla F(x)\| = 2\|a_1 - a_2\|. \tag{84}$$

The denominator is positive and independent of x.

Substituting these identities into the score gives

$$\begin{aligned} u_0(x) &= \frac{-2\alpha\langle a_2 - a_1, x\rangle - \alpha(\|a_1\|^2 - \|a_2\|^2) + b}{2\|a_1 - a_2\|} \\ &= \left\langle \alpha \frac{a_1 - a_2}{\|a_1 - a_2\|}, x \right\rangle + \frac{b - \alpha(\|a_1\|^2 - \|a_2\|^2)}{2\|a_1 - a_2\|}. \end{aligned} \tag{85}$$

This is a linear term plus a bias, with the parameters in Equation equation 82.

Conversely, every Euclidean score $\langle w, x\rangle + c$ with $w \neq 0$ can be represented in this way. Choose $a_1 - a_2$ in the direction of w, set $\alpha = \|w\|$, and take

$$b = 2\|a_1 - a_2\|c + \alpha(\|a_1\|^2 - \|a_2\|^2). \tag{86}$$

For a fixed $\varepsilon > 0$, the denominator remains constant in Euclidean space. The stabilised score is therefore still linear in x with a bias, with

$$\begin{aligned} w_\varepsilon &= \frac{2\alpha(a_1 - a_2)}{2\|a_1 - a_2\| + \varepsilon}, \\ c_\varepsilon &= \frac{b - \alpha(\|a_1\|^2 - \|a_2\|^2)}{2\|a_1 - a_2\| + \varepsilon}. \end{aligned} \tag{87}$$

These identities establish a Euclidean correspondence.

E.6 Proof of Proposition E.5

The Stiefel manifold $\mathrm{St}(n, p)$ and the orthogonal group $O(p)$ are defined in Appendix D.1. We additionally write

$$\mathrm{SO}(p) = \{Q \in O(p) : \det(Q) = 1\} \tag{88}$$

for the special orthogonal group. Here, U denotes a concrete orthonormal basis matrix, while $[U]$ denotes the subspace represented by that basis.

Proposition E.5 (Equivalent orthonormal bases)**.** *Let $U \in \mathrm{St}(n, p)$ and $Q \in O(p)$. Then $\widetilde{U} = UQ$ is an orthonormal basis of the same subspace as U. Conversely, every $V \in \mathrm{St}(n, p)$ satisfying*

$$\mathrm{span}(V) = \mathrm{span}(U) \tag{89}$$

can be written as $V = UQ$ for some $Q \in O(p)$.

Proof. Using the orthogonality constraints,

$$\widetilde{U}^\top \widetilde{U} = Q^\top U^\top U Q = Q^\top Q = I_p. \tag{90}$$

Moreover,

$$\widetilde{U}\widetilde{U}^\top = UQQ^\top U^\top = UU^\top. \tag{91}$$

Thus, the transformed basis remains orthonormal and defines the same orthogonal projector. Its column space is therefore unchanged. Equivalently, since Q is invertible,

$$\mathrm{span}(UQ) = \mathrm{span}(U). \tag{92}$$

Conversely, suppose $V \in \mathrm{St}(n, p)$ spans the same subspace as U. Define

$$Q = U^\top V. \tag{93}$$

Since $UU^\top V = V$,

$$UQ = UU^\top V = V. \tag{94}$$

Furthermore,

$$Q^\top Q = V^\top U U^\top V = V^\top V = I_p. \tag{95}$$

Hence, $Q \in O(p)$, proving the converse.

E.7 Proof of Proposition E.6

For $U \in \mathrm{St}(n,p)$, define its orthogonal projector by

$$P_U = UU^\top. \tag{96}$$

Let $z_1, \dots, z_G \in \mathbb{R}^n$ be learnable probe vectors normalised to unit length. The representation enhancement module defines

$$h_g(U) = z_g^\top P_U z_g, \qquad g = 1, \dots, G. \tag{97}$$

The descriptor vector is

$$h(U) = [h_1(U), \dots, h_G(U)]^\top. \tag{98}$$

Proposition E.6 (Basis invariance of the descriptor)**.** *For every $U \in \mathrm{St}(n,p)$ and every $Q \in O(p)$,*

$$h(UQ) = h(U). \tag{99}$$

Proof. For an equivalent basis UQ,

$$\begin{aligned} P_{UQ} &= (UQ)(UQ)^\top \\ &= UQQ^\top U^\top \\ &= UU^\top \\ &= P_U. \end{aligned} \tag{100}$$

Therefore, for every probe z_g,

$$h_g(UQ) = z_g^\top P_{UQ} z_g = z_g^\top P_U z_g = h_g(U). \tag{101}$$

Stacking these equalities proves $h(UQ) = h(U)$.

The gating network receives $h(U)$ and produces the input-dependent coefficient vector

$$c(U) = \mathrm{MLP}(h(U)) \in \mathbb{R}^G. \tag{102}$$

Proposition E.6 consequently implies

$$c(UQ) = c(U). \tag{103}$$

E.8 Proofs of Propositions E.7 and E.8

Let $a_g, b_g \in \mathbb{R}^p$ be learnable vectors that are normalised before constructing the generators. For each $g = 1, \dots, G$, define

$$\Omega_g = a_g b_g^\top - b_g a_g^\top. \tag{104}$$

Proposition E.7 (Skew-symmetry of the generated Lie-algebra element)**.** *Each Ω_g belongs to the Lie algebra $\mathfrak{so}(p)$. Consequently, the input-dependent linear combination*

$$\Omega_0(U) = \sum_{g=1}^{G} c_g(U)\Omega_g \tag{105}$$

is also an element of $\mathfrak{so}(p)$.

Proof. Transposing Ω_g gives

$$\begin{aligned} \Omega_g^\top &= \left(a_g b_g^\top - b_g a_g^\top\right)^\top \\ &= b_g a_g^\top - a_g b_g^\top \\ &= -\Omega_g. \end{aligned} \tag{106}$$

Thus, Ω_g is skew-symmetric. A real linear combination of skew-symmetric matrices is also skew-symmetric, so

$$\Omega_0(U)^\top = -\Omega_0(U). \tag{107}$$

Therefore, $\Omega_g, \Omega_0(U) \in \mathfrak{so}(p)$.

For a maximum transformation scale $\delta > 0$, define the bounded generator

$$\Omega(U) = \frac{\delta}{\sqrt{\|\Omega_0(U)\|_{\mathrm{F}}^2 + \delta^2}} \Omega_0(U). \tag{108}$$

Proposition E.8 (Bounded generator)**.** *The matrix $\Omega(U)$ is skew-symmetric and satisfies*

$$\|\Omega(U)\|_{\mathrm{F}} < \delta \tag{109}$$

whenever $\delta > 0$.

Proof. The rescaling factor in Equation equation 108 is a nonnegative scalar, so it preserves skew-symmetry. Moreover,

$$\begin{aligned} \|\Omega(U)\|_{\mathrm{F}} &= \frac{\delta \|\Omega_0(U)\|_{\mathrm{F}}}{\sqrt{\|\Omega_0(U)\|_{\mathrm{F}}^2 + \delta^2}} \\ &< \delta. \end{aligned} \tag{110}$$

The strict inequality follows from $\delta > 0$.

E.9 Proof of Proposition 5.1

For the bounded skew-symmetric generator $\Omega(U)$, define

$$R(U) = \left(I_p - \frac{1}{2}\Omega(U)\right)^{-1} \left(I_p + \frac{1}{2}\Omega(U)\right). \tag{111}$$

Proposition E.9 (Special-orthogonal Cayley transformation)**.** *For every real skew-symmetric matrix $\Omega(U)$, the matrix in Equation equation 111 is well defined and satisfies*

$$R(U) \in \mathrm{SO}(p). \tag{112}$$

Proof. All eigenvalues of a real skew-symmetric matrix are either zero or purely imaginary. Therefore, 1 cannot be an eigenvalue of $\frac{1}{2}\Omega(U)$, and

$$I_p - \frac{1}{2}\Omega(U) \tag{113}$$

is invertible. Hence, the Cayley transformation is well defined.

Let

$$H = \frac{1}{2}\Omega(U). \tag{114}$$

Since $H^\top = -H$,

$$\begin{aligned} R(U)^\top &= \left[(I_p - H)^{-1}(I_p + H)\right]^\top \\ &= (I_p - H)(I_p + H)^{-1}. \end{aligned} \tag{115}$$

The matrices $I_p - H$ and $I_p + H$ commute because both are polynomials in H. Consequently,

$$\begin{aligned} R(U)^\top R(U) &= (I_p - H)(I_p + H)^{-1}(I_p - H)^{-1}(I_p + H) \\ &= I_p. \end{aligned} \tag{116}$$

Thus, $R(U) \in O(p)$.

To determine its determinant, consider the continuous path

$$R_t = \left(I_p - \frac{t}{2}\Omega(U)\right)^{-1} \left(I_p + \frac{t}{2}\Omega(U)\right), \qquad t \in [0, 1]. \tag{117}$$

The same eigenvalue argument shows that this path is well defined for every $t \in [0, 1]$. Since each R_t is orthogonal, $\det(R_t) \in \{-1, 1\}$. The determinant is continuous and

$$R_0 = I_p, \qquad \det(R_0) = 1. \tag{118}$$

It follows that $\det(R_t) = 1$ for the entire path, and in particular

$$\det(R(U)) = 1. \tag{119}$$

Therefore, $R(U) \in \mathrm{SO}(p)$.

The enhanced orthonormal basis is defined as

$$\widetilde{U} = UR(U). \tag{120}$$

Proposition E.10 (Orthonormality and subspace preservation)**.** *Let $U \in \mathrm{St}(n,p)$ and let $R(U)$ be defined by Equation equation 111. Then*

$$\widetilde{U} \in \mathrm{St}(n,p) \tag{121}$$

and

$$[\widetilde{U}] = [U]. \tag{122}$$

Equivalently,

$$\widetilde{U}^\top \widetilde{U} = I_p, \qquad \widetilde{U}\widetilde{U}^\top = UU^\top. \tag{123}$$

Proof. By Proposition E.9, $R(U)$ is orthogonal. Therefore,

$$\begin{aligned} \widetilde{U}^\top \widetilde{U} &= R(U)^\top U^\top U R(U) \\ &= R(U)^\top R(U) \\ &= I_p. \end{aligned} \tag{124}$$

Thus, $\widetilde{U} \in \mathrm{St}(n,p)$.

Moreover,

$$\begin{aligned} \widetilde{U}\widetilde{U}^\top &= UR(U)R(U)^\top U^\top \\ &= UU^\top. \end{aligned} \tag{125}$$

Hence, U and $\widetilde{U}$ define the same orthogonal projector and represent the same Grassmann point.

E.10 Proof of Proposition E.11

Proposition E.11 (Quotient-space consistency)**.** *Let $U \in \mathrm{St}(n,p)$ and let $Q \in O(p)$. Suppose that the representation enhancement module constructs its gating coefficients from the projector descriptor in Equation equation 97. Then the enhanced outputs generated from U and UQ represent the same Grassmann point.*

Proof. Proposition E.6 gives

$$c(UQ) = c(U). \tag{126}$$

The generators Ω_g are shared model parameters, and therefore

$$\Omega_0(UQ) = \Omega_0(U). \tag{127}$$

The bounded rescaling and Cayley transformation consequently give

$$\Omega(UQ) = \Omega(U), \qquad R(UQ) = R(U). \tag{128}$$

Let $R = R(U) = R(UQ)$. The two enhanced bases are

$$\widetilde{U} = UR, \qquad \widetilde{U}_Q = UQR. \tag{129}$$

Their projectors satisfy

$$\begin{aligned} \widetilde{U}\widetilde{U}^\top &= URR^\top U^\top \\ &= UU^\top, \end{aligned} \tag{130}$$

and

$$\begin{aligned} \widetilde{U}_Q\widetilde{U}_Q^\top &= UQRR^\top Q^\top U^\top \\ &= UQQ^\top U^\top \\ &= UU^\top. \end{aligned} \tag{131}$$

Therefore,

$$[\widetilde{U}_Q] = [\widetilde{U}] = [U]. \tag{132}$$

The two enhanced bases are also explicitly related by a right orthogonal transformation:

$$\widetilde{U}_Q = \widetilde{U}\left(R^\top QR\right), \tag{133}$$

where $R^\top QR \in O(p)$. Thus, equivalent input bases remain equivalent after representation enhancement.

E.11 Proof of Proposition E.12

Proposition E.12 (Identity initialisation of the enhancement module)**.** *Suppose that the weights and bias of the final linear layer of the gating network are initialised to zero. Then, for every input basis U,*

$$c(U) = 0, \qquad \Omega_0(U) = 0, \qquad \Omega(U) = 0, \qquad R(U) = I_p, \tag{134}$$

and consequently

$$\widetilde{U} = U. \tag{135}$$

Proof. Zero weights and zero bias in the final gating layer imply that its output is the zero vector for every input descriptor:

$$c(U) = 0. \tag{136}$$

Equation equation 105 therefore gives

$$\Omega_0(U) = 0. \tag{137}$$

The bounded rescaling in Equation equation 108 gives

$$\Omega(U) = 0. \tag{138}$$

Substituting this result into the Cayley transformation yields

$$R(U) = (I_p - 0)^{-1}(I_p + 0) = I_p. \tag{139}$$

Hence,

$$\widetilde{U} = UR(U) = U. \tag{140}$$

This identity initialisation ensures that the network begins with the unadjusted Grassmann representation. The representation enhancement module can subsequently learn input-dependent basis transformations through the gating network and the Lie-algebra generators.

F Experiments

F.1 Datasets

FPHA. FPHA contains 1,175 video sequences from 45 first-person hand-action classes (Garcia-Hernando et al., 2018). Each frame records the 3D coordinates of 21 hand joints, resulting in a 63-dimensional temporal feature. For each sequence, we construct a 63×63 covariance matrix from the temporal joint representation. The current implementation uses predefined training and validation partitions. During training, temporal cropping and small Gaussian perturbations are applied to the temporal features, together with shrinkage during covariance construction. The validation samples are processed without random augmentation.

HDM05. HDM05 is a motion-capture dataset for human action recognition (Müller et al., 2007). Following the preprocessing used by previous manifold networks (Huang & Gool, 2017; Huang et al., 2018; Chen et al., 2024a; Yu & Xu, 2026), we use 2,086 sequences from 117 action classes. Each sequence contains the 3D coordinates of 31 joints and is represented by a 93×93 covariance matrix. In the current implementation, all samples are shuffled using a fixed split seed of 42 and equally divided into 1,043 training samples and 1,043 held-out samples. Class-balanced sampling is used during training. For the main comparison, we keep this data partition fixed and train the model ten times from scratch using different random seeds. The final result is reported as the mean accuracy and standard deviation over the ten runs.

Salinas. Salinas is a hyperspectral image acquired by the AVIRIS sensor and contains 16 land-cover classes (Green et al., 1998; Grupo de Inteligencia Computacional). After removing bands affected by water absorption, 204 spectral bands are retained. We first reduce the spectral dimension to 30 using principal component analysis and then construct a 30×30 covariance matrix from a 9×9 neighbourhood around each pixel. For each class, 10% of the samples are used for training, 10% for validation, and the remaining samples for testing. A class-balanced sampler is used during training. Together, the three datasets cover two different data modalities and several Grassmann spaces with different ambient and subspace dimensions. This setting enables us to examine whether the proposed formulation can be used consistently across temporal action representations and local spectral representations.

F.2 Implementation Details

Network. For FPHA and HDM05, the input SPD covariance matrices are first processed by SPD batch normalisation and OrthMap. The resulting subspace representations are subsequently transformed by FRMap and re-orthogonalisation before being passed to GrassMLR. The final representations lie on $\mathcal{G}(56, 20)$ and $\mathcal{G}(70, 20)$ for FPHA and HDM05, respectively. For Salinas, the final representation lies on $\mathcal{G}(30, 10)$.

Comparison Protocol. Following the classifier evaluation strategy used in previous manifold MLR studies (Chen et al., 2024a;b; 2026a), we separate manifold representation learning from the final classification stage. Within each paired comparison, the data preprocessing and feature-extraction network are kept unchanged, while only the final classification module is replaced. We compare GrassMLR with a Euclidean classifier to examine the effect of constructing the final decision directly from Grassmann geometry. We further compare GrassMLR with and without QSRL to examine the effect of quotient-space representation adjustment. For every paired comparison, the same split, preprocessing procedure, feature extractor, and optimisation schedule are retained. This protocol ensures that any difference within a pair is attributable to the classification or representation component being evaluated rather than to a change in the backbone network.

Implementation and Evaluation. All manifold parameters are optimised using Riemannian Adam (Becigneul & Ganea, 2019). Cross-entropy is used as the primary classification objective together with dataset-specific geometric regularisation terms. We report classification accuracy on FPHA and HDM05. For the class-imbalanced Salinas dataset, we report overall accuracy (OA), average accuracy (AA), and the Kappa coefficient. Model size denotes the storage required by the learnable parameters, and FLOPs denote the computational cost of one forward pass under the input configuration of the corresponding dataset. These two quantities are measured using the same counting convention for all implementations evaluated in this work.