INTRINSIC DENSE ASSOCIATIVE MEMORY ON RIEMANNIAN MANIFOLDS

Zhaoyang Shi
Department of Statistics, Havard University
zshi@fas.harvard.edu

Krishna Balasubramanian
Department of Statistics, University of California, Davis
kbala@ucdavis.edu

ABSTRACT

We propose a novel Dense Associative Memory (DenseAM) framework defined intrinsically on a compact Riemannian manifold $\mathcal{M}$, enabling associative memory for manifold-valued data without Euclidean embedding. We introduce two natural geometric extensions of DenseAM on the manifold: (i) Volume-Corrected Geodesic energy (VC-Geodesic energy): a manifold-KDE energy obtained by incorporating the Riemannian volume density correction term, and (ii) Geodesic energy: a purely geodesic energy obtained by removing the correction term. We show that these two formulations exhibit fundamentally different behaviors. The geodesic energy admits exact memorization for finite inverse temperature β, achieves exponential storage capacity in the intrinsic dimension $m = \dim(\mathcal{M})$, and generates abundant emergent memories characterized as local Fréchet means. In contrast, the VC-Geodesic energy introduces a curvature-dependent bias that can destroy exact finite-β memorization, particularly on positively curved manifolds. We further derive intrinsic gradient-based inference dynamics expressed via Riemannian exponential and logarithmic maps, leading to a manifold attention mechanism. Our theories are also supported by preliminary simulations of Riemannian manifold data for statistical inference such as classification and regression.

1 INTRODUCTION

Associative memory models define an energy landscape whose local minima correspond to stored patterns. Beginning with the classical Hopfield network (Hopfield, 1982), energy-based memory systems have been extensively studied in both statistical physics and machine learning. Modern Dense Associative Memory (DenseAM) models (Krotov & Hopfield, 2016; 2018; Demircigil et al., 2017) achieve significantly higher storage capacity by employing nonlinear separation functions. In particular, the exponential-separation formulation called log-sum-exponential (LSE) energy (Ramsauer et al., 2020) establishes a direct connection between DenseAM and Transformer attention. However, a fundamental memorization-generalization trade-off issue appears in the LSE energy based DenseAM. Recent work (Hoover et al., 2025) further demonstrated that a compact-support energy, called log-sum-ReLU (LSR) energy, based on the Epanechnikov kernel can simultaneously achieve exact memorization, exponential capacity, and the emergence of novel local minima. The key discovery of their work lies in revealing the connection of the energy in DenseAM and kernel density estimation (KDE).

Despite these advances, all existing DenseAM formulations are defined in Euclidean spaces. While being theoretically convenient to analyze, many modern datasets naturally reside on nonlinear Riemannian manifolds, including: protein backbone conformations on rotation and shape manifolds (Xu et al., 2020), orthogonal and Stiefel representations in representation learning (Li et al., 2020), Grassmannian subspace models (Murdock & De la Torre, 2017), symmetric positive definite (SPD) matrices in diffusion tensor imaging (Fletcher & Joshi, 2007) and brain connectivity (You & Park, 2021) and covariance geometry in statistics (Arsigny et al., 2007). Riemannian geometry provides the natural mathematical framework for such data. Foundational works in Riemannian statistics include (Pennec, 2006; Bhattacharya & Patrangenaru, 2003; Jacobs, 2014; Guigui et al., 2023). However, energy-based associative memory models have not been developed intrinsically in this geometric setting.

Extending DenseAM to manifolds is nontrivial, particularly given the memorization and generalization trade-off issue. On one hand, simply adopting manifold kernel density estimation as in (Hoover et al., 2025) does not preserve the delicate balance between memorization and generalization established in Euclidean DenseAM. In particular, manifold KDE requires a volume-density correction term arising from the Jacobian of the exponential map. This curvature-dependent factor fundamentally alters the geometry of the energy landscape, thereby preventing exact memorization. In the meantime, we must rigorously characterize the generalization mechanism (i.e., the structure and stability of emergent memories) under manifold geometry, especially given its implications for a manifold Transformer attention design. Hence, a fundamental open problem arises:

Can DenseAM retain exact memorization and meaningful emergence under intrinsic manifold geometry?

In this work, we answer this question affirmatively by developing the first intrinsic dense associative memory framework on a compact Riemannian manifold. We introduce two natural geometric formulations: (i) a **volume- corrected geodesic energy (VC-Geodesic energy**, which incorporates the Riemannian volume density correction term from manifold KDE, and (ii) a **geodesic energy**, defined purely through the geodesic distance. Although both energies are geometrically well-motivated, we show that they exhibit fundamentally different behaviors. The geodesic energy admits exact memorization for finite inverse temperature β, achieves exponential storage capacity in the intrinsic dimension, and generates abundant emergent memories characterized as local Fréchet means. In contrast, the VC-Geodesic energy introduces a curvature-dependent bias that can destroy exact finite-β memorization, particularly on positively curved manifolds such as spheres. We further show that in the large-β regime, the two formulations asymptotically coincide and recover density modes. Thus, generalization (i.e., emergent memories) can be interpreted as high-likelihood regions, effectively capturing the most probable structures or patterns in the manifold-supported data.

Beyond memorization guarantees, we derive intrinsic gradient-based inference dynamics expressed entirely through Riemannian exponential and logarithmic maps, leading to a manifold-native attention mechanism that does not rely on Euclidean embedding.

Contributions. Our main contributions are: (i) The first intrinsic formulation of DenseAM on Riemannian manifolds with exact memorization, exponential capacity scaling in intrinsic manifold dimension while creating a substantial number of new (emergent) memories for generalization. (ii) A mathematically rigorous interpretation of emergent memories as local Fréchet-type means on Riemannian manifolds. (iii) An intrinsic manifold attention mechanism derived from energy gradient dynamics.

2 From Euclidean DenseAM to Manifold Geometry

In Euclidean space, for a set of memories $\Xi = \{\xi_\mu\}_{\mu=1}^M$ in $\mathbb{R}^d$, dense associative memory is defined through an energy of the form

$$E_\beta(x;\Xi) = -\frac{1}{\beta}\log\sum_{\mu=1}^M F(\beta S(x,\xi_\mu)),$$

where S is a similarity function and F is a separation function. When $F(t) = \exp(t)$ and $S(x,\xi) = -\frac{1}{2}\|x-\xi\|^2$, one recovers the log-sum-exponential (LSE) energy, which corresponds to Gaussian kernel density estimation (KDE) as pointed out in (Hoover et al., 2025). While LSE achieves exponential storage capacity, its memorization–generalization trade-off prevents simultaneous exact finite-β retrieval and emergence.

Inspired by the optimal kernel of KDE, the introduction of compact-support kernels, notably the Epanechnikov kernel, leads to the log-sum-ReLU (LSR) energy in (Hoover et al., 2025),

$$E_\beta^{\mathrm{LSR}}(x) = -\frac{1}{\beta}\log\left(\varepsilon + \sum_{\mu=1}^M \left[1-\frac{\beta}{2}\|x-\xi_\mu\|^2\right]_+\right),$$

which preserves exact finite-β memorization while generating abundant emergent memories. The key structural insight is that DenseAM energies are intrinsically connected to kernel density estimation.

Using the KDE perspective, on a Riemannian manifold $\mathcal{M}$ with dimension m and geodesic distance d_g, kernel smoothing must account for curvature-induced volume distortion. Let $\mathrm{d}V_g$ denote the Riemannian volume measure and let $\exp_p : T_p\mathcal{M} \to \mathcal{M}$ be the exponential map. For $q \in \mathcal{M}$ with $d_g(p, q) < \mathrm{inj}(\mathcal{M})$, write $v = \log_p(q) \in T_p\mathcal{M}$ so that $q = \exp_p(v)$. The *volume density function* (or *Jacobian factor*) is defined by

$$\theta_p(q) := \left|\det\big(D\exp_p(v)\big)\right|, \qquad v = \log_p(q),$$

equivalently the change-of-variables identity

$$\mathrm{d}V_g(q) = \theta_p(q)\,\mathrm{d}v \quad \text{under} \quad q = \exp_p(v),$$

where $\mathrm{d}v$ is Lebesgue measure on $T_p\mathcal{M} \cong \mathbb{R}^m$ induced by the metric at p. Let $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a radial kernel with compact support, and define the bandwidth-scaled kernel

$$K_h(r) := \frac{1}{h^m} K\left(\frac{r}{h}\right), \qquad h > 0.$$

Given an i.i.d. sample $\{\xi_\mu\}_{\mu=1}^M \subset \mathcal{M}$ from a density p supported on $\mathcal{M}$, the intrinsic manifold KDE is

$$\widehat{p}_h(x) := \frac{1}{M} \sum_{\mu=1}^{M} \frac{1}{\theta_{\xi_\mu}(x)}\, K_h\big(d_g(x, \xi_\mu)\big),$$

where the correction factor $1/\theta_{\xi_\mu}(x)$ compensates for the non-Euclidean volume element in geodesic normal coordinates.

3 Two Possible Manifold Energies

The manifold KDE naturally inspires us to consider two intrinsic DenseAM formulations as follows.

The *Volume-Corrected Geodesic Energy*

$$E_\beta^{\mathrm{vc}}(x) = -\frac{1}{\beta} \log\left(\varepsilon + \sum_{\mu=1}^{M} \frac{1}{\theta_{\xi_\mu}(x)} \left[1 - \frac{\beta}{2} d_g(x, \xi_\mu)^2\right]_+\right),$$

incorporates the metric–measure structure.

The *Geodesic Energy*

$$E_\beta^{\mathrm{geo}}(x) = -\frac{1}{\beta} \log\left(\varepsilon + \sum_{\mu=1}^{M} \left[1 - \frac{\beta}{2} d_g(x, \xi_\mu)^2\right]_+\right),$$

depends solely on the Riemannian metric.

The difference between the two formulations is subtle but profound. Locally, if $q = \exp_p(v)$ with $\|v\| \ll 1$,

$$\theta_p(q) = 1 - \frac{1}{6}\mathrm{Ric}_p(v, v) + o(\|v\|^2),$$

so curvature enters explicitly. On positively curved manifolds such as the sphere S^m, $\theta_p(q)$ is maximized at $q = p$. Consequently, $1/\theta_p(q)$ is minimized at p, and the volume-corrected energy introduces a curvature-dependent bias term. When the active set contains a single memory ξ_μ, the effective objective becomes

$$F_\beta(x) = \frac{1}{\theta_{\xi_\mu}(x)} d_g(x, \xi_\mu)^2 - \frac{2}{\beta} \frac{1}{\theta_{\xi_\mu}(x)}.$$

Although $x = \xi_\mu$ minimizes the squared distance, it does not necessarily minimize the Jacobian term. Therefore, for finite β, stored memories may cease to be strict local minima. Exact memorization is generally recovered only in the limit $\beta \to \infty$.

This illustrates that incorporating the volume correction fundamentally reshapes the energy landscape under curvature thus breaking the simultaneous memorization-generalization ability as demonstrated in the Euclidean case in (Hoover et al., 2025). Consequently, the geodesic energy E_β^{geo} is preferred.

4 PROPERTIES OF THE GEODESIC ENERGY

Let $(\mathcal{M}, g)$ be a compact m-dimensional Riemannian manifold with injectivity radius $\mathrm{inj}(\mathcal{M}) > 0$. Let $\Xi = \{\xi_\mu\}_{\mu=1}^M \subset \mathcal{M}$ denote stored memories.

4.1 EXACT MEMORIZATION AND ONE-STEP RETRIEVAL

Define the minimum pairwise separation $r := \min_{\mu \neq \nu} d_g(\xi_\mu, \xi_\nu)$, and choose $\Delta \in (0, r) \cap (0, \mathrm{inj}(\mathcal{M}))$. Set $\beta = \frac{2}{(r-\Delta)^2}$. For any $x \in B_g(\xi_\mu, \Delta)$, the gradient of the Geodesic Energy takes the form

$$\nabla E_\beta^{\mathrm{geo}}(x) = \frac{\log_x(\xi_\mu)}{\varepsilon + \left(1 - \frac{\beta}{2} d_g(x, \xi_\mu)^2\right)}.$$

Consequently: we have the following:

1. Each stored memory ξ_μ is a strict local minimum.
2. Riemannian gradient descent initialized in $B_g(\xi_\mu, \Delta)$ converges to ξ_μ.
3. Define $S(x) = \varepsilon + \left(1 - \frac{\beta}{2} d_g(x, \xi_\mu)^2\right)$. Inside $B_g(\xi_\mu, \Delta)$, the gradient satisfies $\nabla E_\beta^{\mathrm{geo}}(x) = -\frac{\log_x(\xi_\mu)}{S(x)}$. Choosing the step size $\eta = S(x)$ yields

   $$x^+ = \exp_x(-\eta \nabla E_\beta^{\mathrm{geo}}(x)) = \exp_x(\log_x(\xi_\mu)) = \xi_\mu,$$

   establishing one-step retrieval.

Thus exact finite-β memorization is preserved intrinsically.

4.2 EXPONENTIAL CAPACITY

We impose the following standard geometric and sampling assumptions:

1. $\mathcal{M}$ is a compact m-dimensional Riemannian manifold with injectivity radius $\mathrm{inj}(\mathcal{M}) > 0$ and finite volume $V = \mathrm{vol}(\mathcal{M})$.
2. The stored memories $\{\xi_\mu\}_{\mu=1}^M$ are drawn i.i.d. from a distribution on $\mathcal{M}$ whose density is bounded above and below by positive constants.
3. The inverse temperature β is chosen so that the induced geodesic kernel radius $r_\beta = \sqrt{2/\beta}$ is strictly smaller than the injectivity radius.

Under these assumptions, the geodesic energy achieves exponential storage capacity in intrinsic dimension. Specifically, for any confidence level $1 - \delta$, there exists $\alpha > 0$ such that

$$M = \Theta\left(e^{\alpha m} \sqrt{1 - \delta}\right)$$

stored memories can be embedded with high probability while preserving exact finite-β retrieval and distinct basins of attraction.

4.3 EMERGENT MEMORIES

In the above sections, we have demonstrated the desired memorization properties of the geodesic energy. In the section, we mathematically characterize the generalization of the geodesic energy, i.e., its emergent memories (see definition in (Hoover et al., 2025)).

Let the activation set for any point $x \in \mathbb{R}^d$ be $B(x) = \left\{\mu : \ d_g(x, \xi_\mu)^2 \leq \frac{2}{\beta}\right\}$. Critical points of E_β^{geo} satisfy

$$\sum_{\mu \in B(x)} \log_x(\xi_\mu) = 0.$$

Therefore, emergent memories are precisely local Fréchet means of active subsets of stored memories. When $|B(x)| = 1$, the solution corresponds to a stored memory. When $|B(x)| > 1$, the solution is a novel memory. Geodesic strong convexity within fixed-active regions ensures basin stability of such critical points.

4.4 Counting Emergent Memories

Let $r_\beta = \sqrt{2/\beta}$ and $V = \mathrm{vol}(\mathcal{M})$. Due to i.i.d. sampling, by concentration of measure, the expected active set size scales as $|B(x)| \asymp M V_m r_\beta^m$, where V_m is the volume of the unit ball in $\mathbb{R}^m$. Hence, a combinatorial counting argument yields the number of such emergent memories is

$$O\left(\exp\left(V V_m r_\beta^m \log \frac{e}{V_m r_\beta^m}\right)\right),$$

showing that the number of emergent memories can grow superlinearly in M while exact memorization is retained.

5 Intrinsic Manifold Attention

The update of the LSE energy in DenseAM has been shown its equivalence to the attention in Transformer (Ramsauer et al., 2020). This underpins a mathematical connection between the update of DenseAM to attention mechanism. Hence, inspired by this connection and the Fréchet mean update, we propose a novel manifold attention based on LSE energy.

Given a query $q \in \mathcal{M}$, keys $\{k_\mu\}_{\mu=1}^M \subset \mathcal{M}$, and values $\{v_\mu\}_{\mu=1}^M \subset \mathbb{R}^d$, we define intrinsic manifold softmax attention weights by

$$\alpha_\mu^{\mathcal{M}}(q) := \frac{\exp\left(-\frac{\beta}{2} d_g(q, k_\mu)^2\right)}{\sum_{\nu=1}^M \exp\left(-\frac{\beta}{2} d_g(q, k_\nu)^2\right)}.$$

The intrinsic manifold attention output is

$$q' := \exp_q\left(\sum_{\mu=1}^M \alpha_\mu^{\mathcal{M}}(q) \log_q(k_\mu)\right).$$

This defines a fully intrinsic attention mechanism without Euclidean embeddings: similarity is measured by geodesic distance, aggregation is performed in the tangent space $T_q\mathcal{M}$, and the exponential map returns the updated query to $\mathcal{M}$. In Section in the appendix, we apply this manifold attention for statistical inference problems such as classification and regression to evaluate its numerical performance.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

P Thomas Fletcher and Sarang Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

Nicolas Guigui, Nina Miolane, Xavier Pennec, et al. Introduction to riemannian geometry and geometric statistics: from basic theory to implementation with geomstats. *Foundations and Trends® in Machine Learning*, 16(3):329–493, 2023.

Benjamin Hoover, Zhaoyang Shi, Krishnakumar Balasubramanian, Dmitry Krotov, and Parikshit Ram. Dense associative memory with epanechnikov energy. *arXiv preprint arXiv:2506.10801*, 2025.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Justin Wayne Jacobs. *Nonparametric density estimation on riemannian manifolds*. University of Maryland, Baltimore County, 2014.

Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12):3151–3167, 2018.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

Calvin Murdock and Fernando De la Torre. Approximate grassmannian intersections: Subspace-valued subspace learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4308–4316, 2017.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Gang Xu, Qinghua Wang, and Jianpeng Ma. Opus-tass: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, 36(20):5021–5026, 2020.

Kisung You and Hae-Jeong Park. Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:117464, 2021.

A NUMERICAL EVALUATION: CLASSIFICATION ON S^2

We evaluate the proposed intrinsic manifold attention on a synthetic classification task defined on the unit sphere S^2. We compare the performance of manifold attention we propose with the classic Euclidean gradient descent followed by projection.

We sample $K = 3$ class centers $\{c_k\}_{k=1}^{K} \subset S^2$ with minimum angular separation 30°. For each class k, we generate $n_{\text{key}} = 20$ keys by sampling on the geodesic ball $B_g(c_k, r_{\text{key}})$ with radius $r_{\text{key}} = 0.35$ (radians), yielding $M = K\, n_{\text{key}} = 60$ keys in total. Queries are sampled more broadly from $B_g(c_k, r_{\text{qry}})$ with $r_{\text{qry}} = 0.75$, with $n_{\text{qry}} = 100$ queries per class ($N_{\text{test}} = 300$ total).

For each query $q \in S^2$, we apply $T = 5$ iterative update layers parameterized by inverse bandwidth β. We compare: (i) an intrinsic update on S^2 using logarithmic and exponential maps (our proposed manifold attention), and (ii) an extrinsic Euclidean attention update in $\mathbb{R}^3$ followed by projection $x \mapsto x/\|x\|$ onto S^2. We sweep β over a log-grid $\beta \in \{0.1, \ldots, 100\}$ (20 values), use step size $\text{lr} = 0.30$ per layer for both methods, and average results over $S = 5$ random seeds. Performance is reported as the mean geodesic distance $d_g(\hat{q}_T, c_y)$ between the final iterate and the true class center.

Figure 1 shows the mean geodesic error as a function of β. The manifold attention inference consistently achieves lower intrinsic distance to the correct center across a wide range of β values compared to the classic Euclidean attention inference. The gap becomes pronounced in regimes where curvature effects are significant, highlighting the instability introduced by Euclidean updates under nonlinear geometry. Notably, the manifold attention inference exhibits a clear optimal β region, corresponding to well-balanced local interaction, while Euclidean attention updates degrade under either over-smoothing or excessive sparsity.

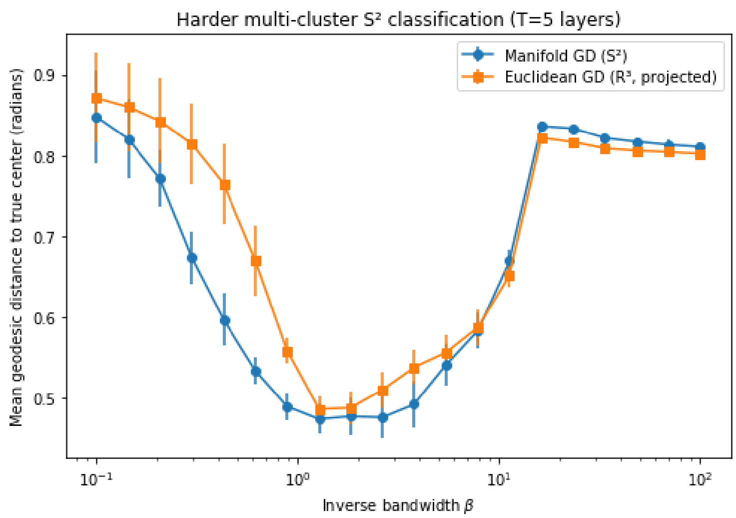

Figure 1: Mean geodesic error of manifold attention and Euclidean attention in classification tasks on S^2.

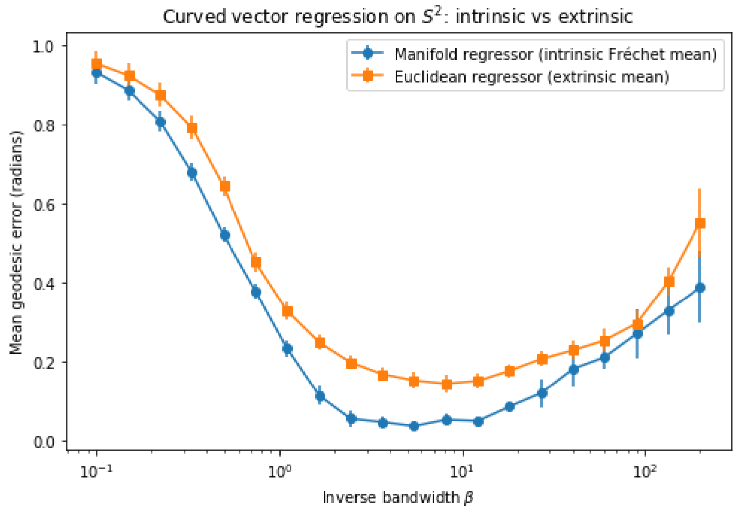

Figure 2: Mean geodesic error of manifold attention and Euclidean attention in regression tasks on S^2.

B Numerical Evaluation: Regression on S^2

We next evaluate manifold attention versus classic Euclidean attention plus projection on regression tasks on S^2 under a deliberately curvature-sensitive target map. We define a piecewise-rotational function $f : S^2 \to S^2$ that applies a large rotation about a fixed axis, with the rotation direction determined by the input longitude. This produces multi-modal outputs and makes Euclidean averaging unreliable.

Training inputs are sampled uniformly from a spherical cap $\{x \in S^2 : \; x_3 \geq \cos(90°)\}$. We use $N_{\text{train}} = 500$ training points and $N_{\text{test}} = 10$ test points per trial. Noisy outputs are generated as

$$y_i = \exp_{f(x_i)}(\sigma u_i), \qquad \sigma = 0.6,$$

where $u_i \in T_{f(x_i)}S^2$ is a random unit tangent direction. We compare: (i) a Euclidean attention based regressor given by the normalized Euclidean mean $\hat{y}_{\text{ext}}(x) \propto \sum_i w_i(x) y_i$, where distances for the weights are Euclidean in $\mathbb{R}^3$ with projection after the final update; and (ii) a manifold attention based regressor given by the weighted Fréchet mean

$$\hat{y}_{\text{int}}(x) = \arg\min_{y \in S^2} \sum_i w_i(x)\, d_g(y, y_i)^2.$$

We sweep β over a log-grid $\beta \in \{0.1, \ldots, 200\}$ (20 values), average over $S = 5$ random seeds, and report mean geodesic error $\frac{1}{N_{\text{test}}} \sum_j d_g(\hat{y}(x_j), f(x_j))$ with standard error.

As shown in Figure 2, manifold attention based intrinsic regression achieves uniformly smaller geodesic error, with the largest gains occurring in intermediate β regimes where local averaging must remain curvature-consistent. This supports the view that manifold Fréchet-type aggregation is essential for manifold-valued prediction and cannot be replicated by Euclidean averaging followed by projection.