Curvature Estimation on Data Manifolds with Diffusion-Augmented Sampling

Jason Wang, Bobak Kiani, Melanie Weber

Introduction

Often, the first step to analyze a data manifold is to estimate one number: its intrinsic dimension.

What if we want a more complete picture?

Curvature is the fundamental descriptor of local geometry—useful in shape analysis, learning theory, and non-Euclidean algorithms—yet it proves elusive to estimate on sparse, noisy data.

Notation

Manifold $\mathcal{M}$ with intrinsic dimension d and normal dimension n. We denote $T_p\mathcal{M}$ as the tangent space and $\frac{\partial}{\partial x^i}$ as a tangent basis vector.

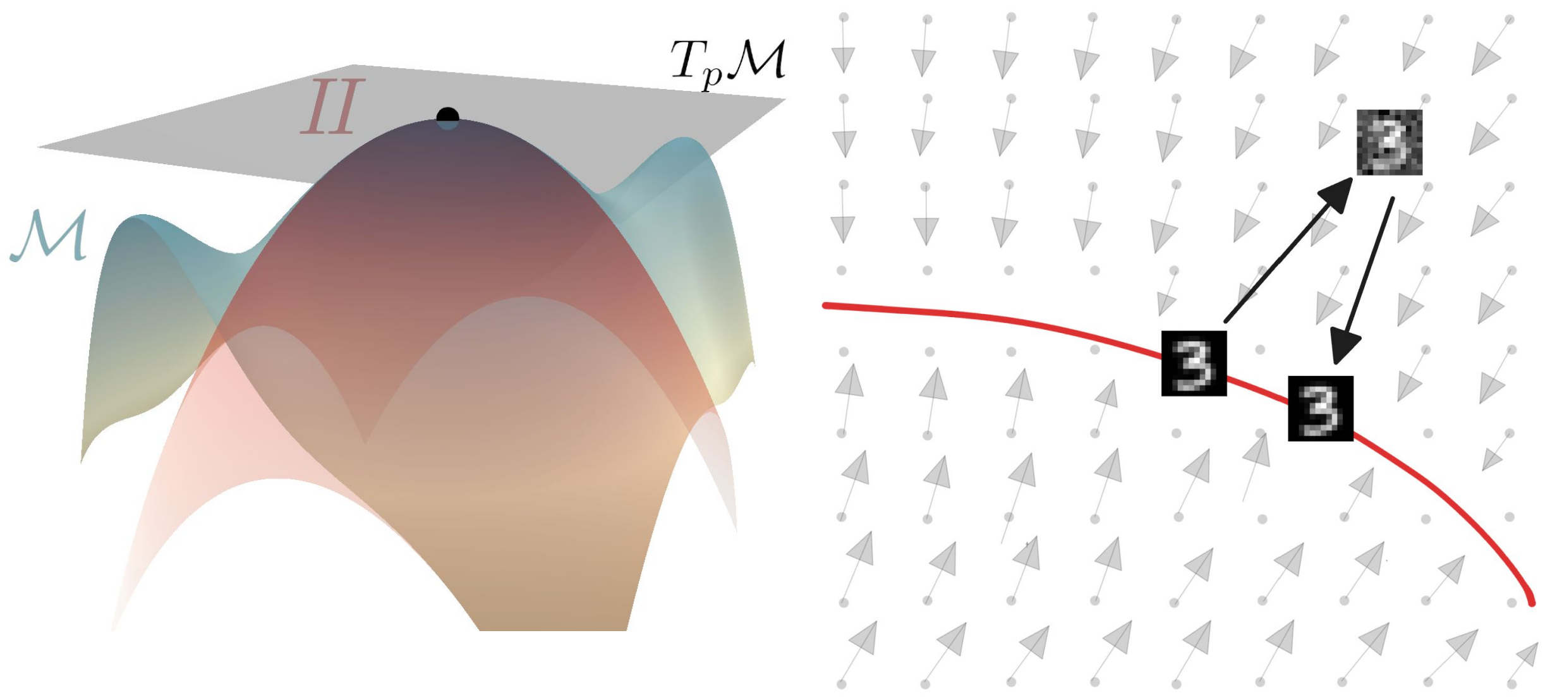
Second Fundamental Form (SFF)

The most complete curvature notion: a $d \times d \times n$ tensor; the Hessian of the manifold viewed from the tangent space for each normal direction.

$$\mathrm{II}_p : T_p\mathcal{M} \times T_p\mathcal{M} \to N_p\mathcal{M},\ h^k_{ij} = \langle \frac{\partial \phi}{\partial x^i \partial x^j}, n^k \rangle$$

Normal Curvature

Curvature of a single curve on the manifold's surface: $\kappa_N = ||\mathrm{II}_p(\gamma', \gamma')||$



Diffusion Model

A generative model that learns to denoise a noised sample, which corresponds to score matching. The score points to the data manifold.

Contributions

Task: Estimate Curvature from Manifold
Challenge: Sensitivity to Sample Quality, Curse of Dimensionality
Contributions:

1. Propose **training a diffusion model** to augment the given data samples.
2. Introduce a **novel estimator for normal curvature** using diffusion-generated geodesic paths.
3. **Optimize neighborhood selection** with streaming algorithms.
4. Investigate the efficacy of various curvature estimators **under varying data conditions and on real world manifolds.**

Methods

Quadratic Regression (Cao et al. 2021)

1. Over locally linear neighborhood, perform a PCA to obtain tangent basis.
2. Over locally quadratic neighborhood, regress the manifold using degree 2 terms of the tangent basis. The result is the SFF.

Diffusion Maps (Jones 2024)

1. Compute the normalized graph Laplacian of the heat kernel applied to the dataset. This is an approximation of the Laplace-Beltrami operator.
2. Invoke iterated Carre du Champ identity to get the SFF from the Laplacian.

Diffusion-Augmented Regression (Ours)

1. Train a diffusion model to learn the data manifold.
2. Use the diffusion model to generate a locally linear neighborhood and perform PCA (thereby learning the intrinsic dimension).
3. Use the diffusion model to generate a locally quadratic neighborhood and perform quadratic regression to compute SFF.

Geodesic Interpolation (Ours)

1. Train a diffusion model to learn the data manifold.
2. Specify a direction of interest, and linearly interpolate along that direction.
3. Use a diffusion model to denoise the linear interpolants back onto the manifold.
4. Compute Frenet-Serret frame via Gram-Schmidt orthogonalization on successive derivatives, from which normal curvature may be computed.

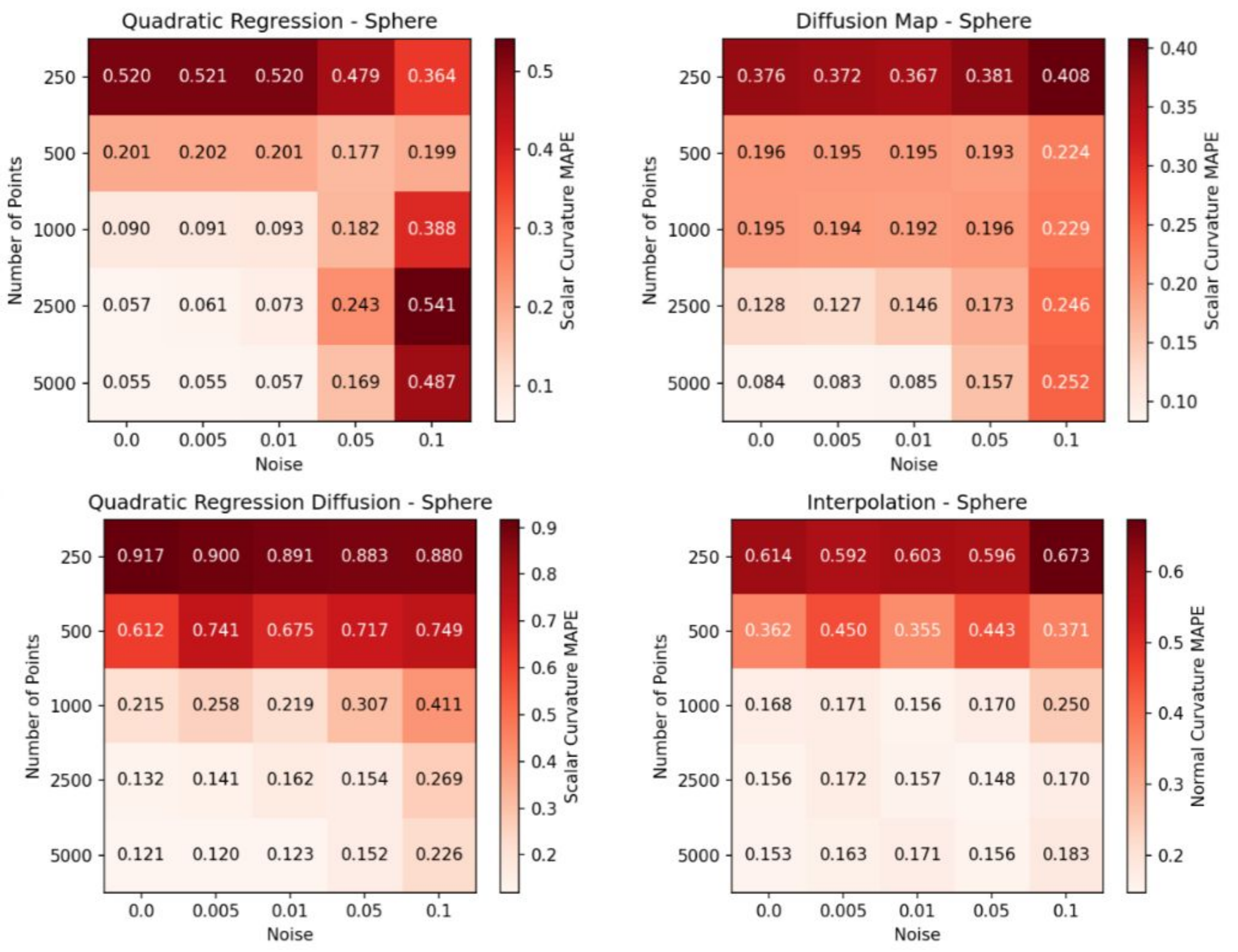
Neighborhood Selection via Streaming Algorithms (Ours)

Local linear and local quadratic neighborhoods are sensitive hyperparmeters to tune. Instead of brute force search, we use streaming algorithms (Incremental PCA, Recursive Least Squares) to try incrementally larger neighborhoods.

Experiments

Noise vs. Sample Density (Sphere and More)

- Sample Count: [250, 500, 1000, 2000, 5000]
- Gaussian Noise Var: [0, 0.005, 0.01, 0.05, 0.1]



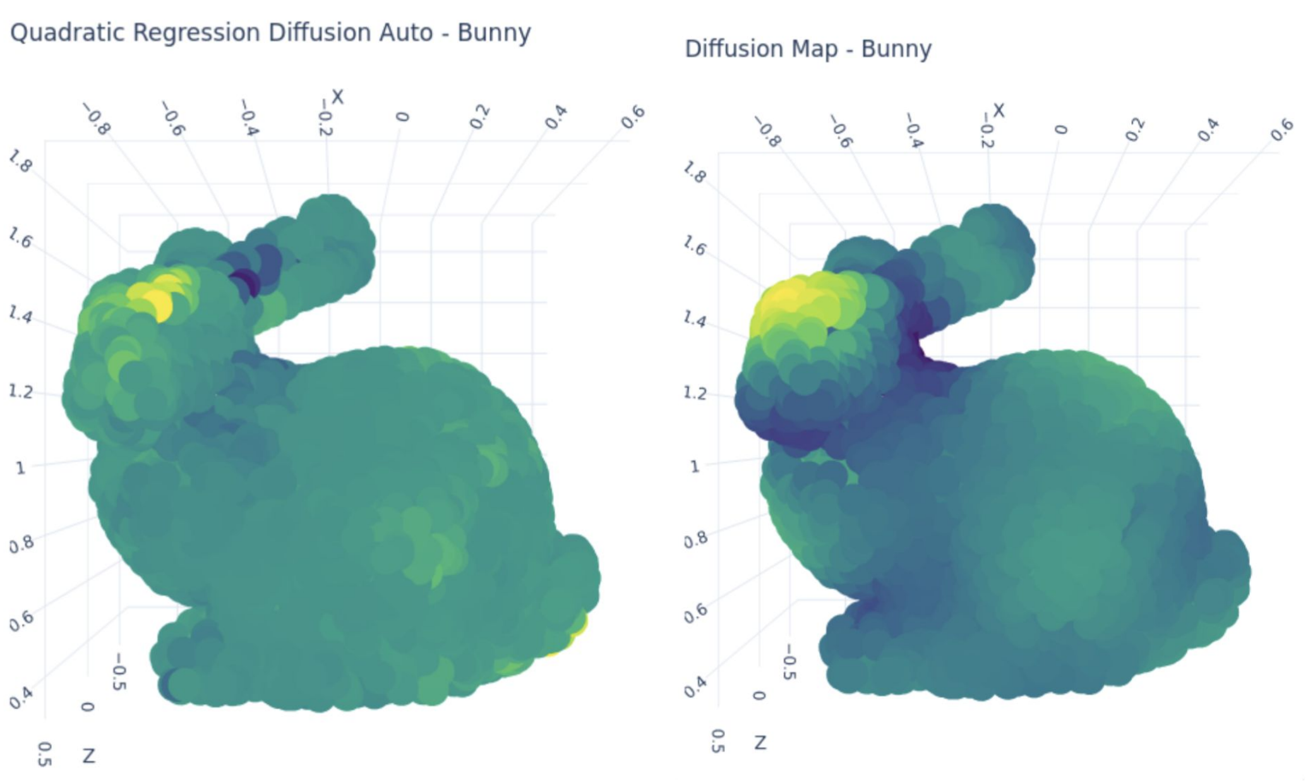
- If you have sufficient samples and little noise, quadratic regression is sufficient
- But diffusion-based sampling is the preferred algorithm for noisy samples

Intrinsic vs. Normal Dim. (Paraboloid)

- Intrinsic, Normal Dim.: [2, 4, 8, 16, 32]
- Quadratic regression and geodesic interpolation scale relatively well

Real World Data Manifold (Bunny)

- Diffusion maps perform the best as it leverages global smoothing, but our diffusion approach is the next closest



Conclusion

Method	Complexity
Quad. Regression	$O(Nd^2n)$
Diffusion Maps	$O(N^3)$
Geod. Interpolation	$O(N(d+n))$

- Diffusion models yield denser samples at the cost of higher noise; this trade-off is particularly good for heterogeneous manifolds.
- Geodesic interpolation scales well with dimension.
- Future work points to leveraging more powerful diffusion models and higher dimensional real world datasets.