Exploiting Data Geometry in Machine Learning

Melanie Weber

Harvard University
mweber@seas.harvard.edu

Abstract

A key challenge in Machine Learning (ML) is the identification of geometric structure in high-dimensional data. Most algorithms assume that data lives in a high-dimensional vector space; however, many applications involve non-Euclidean data, such as graphs, strings and matrices, or data whose structure is determined by symmetries in the underlying system. Here, we discuss methods for identifying geometric structure in data and how leveraging data geometry can give rise to efficient ML algorithms with provable guarantees.

Identifying Geometric Structure in Data

High-dimensional data with intrinsic geometric structure can often be represented with high accuracy in a low-dimensional non-Euclidean space. This has motivated the development of algorithms for identifying representation spaces whose geometric structure aligns with that of the data (Weber 2020; Trillos and Weber 2023). Reparametrizing a function on a suitable geometric domain can give rise to more efficient optimization subroutines. A prime example are Euclidean nonconvex matrix-valued problems, which fulfill a Riemannian notion of convexity with respect to a suitable non-Euclidean metric (Weber and Sra 2022). Examples include common subroutines in ML algorithms, such as barycenter problems, robust subspace recovery, maximum likelihood estimation and the computation of generalized eigenvalues.

Geometric Methods for Graph Machine Learning

Graph-structured data is ubiquitous in the Sciences and Engineering; Graph ML has had a tremendous impact on scientific discovery, biomedical research and in the study of social networks, among others. Graphs may be viewed as geometric objects, a perspective that allows for translating key concepts, such as curvature, from Differential Geometry to the discrete setting. Several discrete analogues of Ricci curvature have proven useful for Graph ML (Weber, Saucan, and Jost 2017). Curvature-based algorithms for unsupervised node clustering have been shown to be more efficient and scalable than related approaches based on the spectrum of the Graph Laplacian or modularity optimization (Tian, Lubberts, and Weber 2023). Moreover, curvature can be used to improve the performance of Graph Neural Networks, including via curvature-based rewiring (Fesser and Weber 2023b) and curvature-based structural encodings (Fesser and Weber 2023a).

Machine Learning on Manifolds

Encoding data geometry as inductive bias into ML architectures can often lead to algorithmic benefits.We will review several results on representation trade-offs in ML algorithms with and without geometric inductive biases. (Weber and Sra 2022, 2023) show that exploiting data geometry in nonconvex optimization allows for certifying fast convergence of first-order methods, a result which applies to a range of classical ML tasks. (Weber et al. 2020) investigates the role of data geometry in robust classification of data with hierarchical structure and demonstrates the advantages of geometric inductive biases both theoretically and computationally.

References

Fesser, L.; and Weber, M. 2023a. Effective Structural Encodings via Local Curvature Profiles. *arXiv:2311.14864*.

Fesser, L.; and Weber, M. 2023b. Mitigating Over-Smoothing and Over-Squashing using Augmentations of Forman-Ricci Curvature. *Learning on Graphs Conference*.

Tian, Y.; Lubberts, Z.; and Weber, M. 2023. Curvature-based clustering on graphs. *arXiv:2307.10155*.

Trillos, N. G.; and Weber, M. 2023. Continuum Limits of Ollivier's Ricci Curvature on data clouds: pointwise consistency and global lower bounds. *arXiv preprint arXiv:2307.02378*.

Weber, M. 2020. Neighborhood Growth Determines Geometric Priors for Relational Representation Learning. In *International Conference on Artificial Intelligence and Statistics*, volume 108, 266–276.

Weber, M.; Saucan, E.; and Jost, J. 2017. Characterizing Complex Networks with Forman-Ricci Curvature and Associated Geometric Flows. *Journal of Complex Networks*, 5(4): 527–550.

Weber, M.; and Sra, S. 2022. Riemannian Optimization via Frank-Wolfe Methods. *Mathematical Programming*.

Weber, M.; and Sra, S. 2023. Global optimality for Euclidean CCCP under Riemannian convexity. In *International Conference on Machine Learning*.

Weber, M.; Zaheer, M.; Rawat, A. S.; Menon, A.; and Kumar, S. 2020. Robust Large-Margin Learning in Hyperbolic Space. In *Advances in Neural Information Processing Systems 34*.