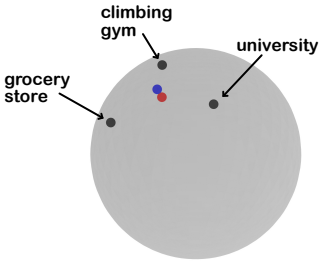
Yangdi Jiang | **Research** Statement

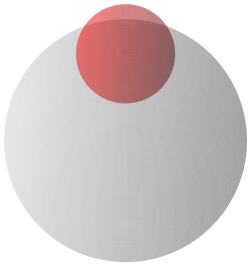
Current research

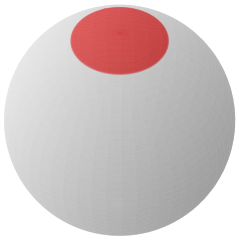
Suppose you are searching for a new apartment and need to commute daily to the university, a grocery store, and a climbing gym. Your goal is to minimize the total commuting cost, which is assumed to be proportional to the square of the total distances travelled. A common approach is to treat each location as a point in three-dimensional Euclidean space and compute their average location. However, since we live on a spherical Earth, this method would incorrectly place your ideal apartment *beneath the ground*, indicated by the red point in the figure. To avoid such nonsensical results, we need the tools of differential geometry, which will ensure the new location remains on the sphere, indicated by the blue point in the figure. Computing a geometric mean in this framework is a classic example of ***performing statistics on unconventional spaces***. As technological advancements accelerate, we increasingly face the challenge of analyzing complex data residing in nonlinear manifolds, which arise in fields such as medical imaging [1, 2, 3], signal processing [4, 5], computer vision [6, 7, 8], and geometric deep learning [9, 10]. These manifolds have distinct geometric properties that can be leveraged to extract meaningful insights, but conventional statistical tools are insufficient for their analysis. Just as choosing an apartment based on Euclidean averaging can lead to unrealistic results, analyzing complex data without accounting for its underlying geometry can lead to incorrect conclusions, making differential geometry an essential tool for modern statistical inference.



new apartment location

Returning to the apartment problem, suppose an evil third party wants to locate your new home for malicious purposes. They know your selection criteria (minimizing commute distance) and the three locations you frequent. With some knowledge of differential geometry, they could pinpoint your apartment with precision. To evade them, you shouldn't settle exactly at the geometric average but instead choose a location randomly within its proximity. But what should this proximity be? The conventional approach treats points on the sphere as three-dimensional coordinates in Euclidean space, leading to a three-dimensional proximity region. However, randomly selecting a point within this region introduces a nonzero probability of ending up in *midair* or *underground*. To avoid such nonsensical results, we again turn to differential geometry, which considers points intrinsically as part of a two-dimensional manifold rather than embedding them in three-dimensional space. Under this framework, the proximity remains entirely on the sphere, ensuring your new apartment stays on solid ground. This simple example illustrates a broader issue: ***maintaining privacy when dealing with data on manifolds***.


3-d proximity region


correct 2-d proximity region

As data becomes increasingly complex, the task of safeguarding privacy also becomes more challenging and intricate. Differential Privacy (DP) [11], a leading mathematical framework, has been widely recognized for its ability to quantify and ensure privacy protection. While numerous mechanisms have been developed to achieve DP [12, 13, 14, 15], these traditional mechanisms, primarily designed for linear data, often fall short when applied to complex nonlinear data (as illuminated in the apartment example). For instance, the commonly adopted extrinsic method embeds nonlinear data into Euclidean space, allowing the application of standard DP mechanisms. However, it has been demonstrated that leveraging the intrinsic properties of nonlinear data can significantly enhance data utility while maintaining privacy [16]. This underscores the necessity for privacy mechanisms that integrate differential geometry to effectively address the complexities of nonlinear data and fully leverage its geometric structure.

Currently, my work centers around developing privacy-preserving mechanisms on Riemannian manifolds. The core of the differential privacy research consists of achieving a better trade-off between data utility and privacy. On the one hand, data sharing is a key part of scientific research, and the wide availability of data reduces the friction of conducting research. Furthermore, good researches are rarely done on bad data. Thus, it is important to have good-quality data. On the other hand, there is the concern of data privacy as some third party could use this data with ill intentions. A growth of one aspect would cause the decline of the other. A high-quality privacy-preserving mechanism should achieve good data utility (that is, good data quality) while maintaining the same amount of privacy protection. Building on my previous works on privacy [17, 18], I am able to develop high-quality privacy-preserving mechanisms by exploring the intrinsic geometry properties of these complex data using the tools from differential geometry [19]. In particular, I extend an important variation of the differential privacy definition, Gaussian differential privacy, to general Riemannian manifolds, develop a novel Riemannian Gaussian mechanism and demonstrate its superior performance compared to the previously proposed mechanism. Building on this work, I extend the framework to Hadamard manifolds, a class of non-positively curved spaces [20]. Leveraging the Cartan–Hadamard theorem, I introduce Exponential-Wrapped Laplace and Gaussian mechanisms that achieve ε-DP, (ε, δ)-DP, Gaussian DP, and Rényi DP in these settings. Unlike traditional Monte Carlo Markov Chain (MCMC) methods, this approach employs computationally efficient, easily implementable algorithms. This work is the first to extend (ε, δ)-DP, GDP, and RDP to Hadamard manifolds. Extensive simulations and real-world applications demonstrate that this approach significantly improves utility at smaller privacy budgets and runs orders of magnitude faster than existing Riemannian mechanisms, reinforcing their practical relevance in privacy-preserving data analysis.

Future research agenda

My research lies at the intersection of differential geometry and statistics, driven by the increasing need to analyze complex data that resides in nonlinear spaces. As data collection technology advances, such data are becoming more prevalent in fields like medical imaging, robotics, and neuroscience, where traditional Euclidean methods often fail to capture the underlying structure. To extract meaningful insights, it is essential to develop statistical methodologies that incorporate the unique geometric properties of this data. My work has deepened my interest in understanding how geometry shapes statistical methods, revealing several important research directions. These include developing differentially private Bayesian nonparametric frameworks on manifolds, constructing differentially private manifold learning techniques, incorporating topological data analysis (TDA) into Geometric Statistics, and extending post-selection inference to general manifolds. Advancing these areas will not only strengthen the theoretical foundations of geometric statistics but also drive impactful applications in data privacy, machine learning, and scientific discovery. Below, I outline these directions in more detail.

Develop DP Bayesian Nonparametric Framework on General Manifold. Statistical inference is the cornerstone of data analysis, enabling researchers to draw meaningful conclusions and make informed decisions. However, methods for performing statistical inference under DP on manifolds remain scarce and underdeveloped. To bridge this gap, I aim to develop a DP Bayesian nonparametric framework specifically designed for manifold data. These methods are highly adaptable, offering flexible modelling without imposing strict assumptions about the data, and are particularly suited for complex or uncertain situations. Unlike frequentist approaches, Bayesian methods naturally incorporate prior knowledge, account for uncertainty, provide full probability distributions for predictions, and mitigate overfitting through probabilistic regularization. However, applying Bayesian nonparametric methods to manifolds presents unique challenges due to the complexity of defining priors and performing computationally efficient inference on these structures. Incorporating DP constraints adds further complexity, as it requires safeguarding sensitive data while preserving statistical utility. Building on the previous connection established between differential privacy mechanisms and measurement error models [18], the proposed framework will integrate the DP mechanisms we previously developed [19] into a Bayesian nonparametric context, enabling tasks such as Fréchet mean estimation and density estimation on manifolds while maintaining data privacy. By addressing these challenges, this work fills a critical gap in the literature, offering a powerful approach for analyzing manifold data privately and effectively.

Differentially Private Manifold Learning. Another promising direction is performing statistical tasks on unknown manifolds. My work so far has assumed that the underlying manifold is known, allowing us to leverage its geometry for inference. However, when the manifold is unknown, the problem shifts dramatically, resembling dimension reduction. Real-world data—such as images, sounds, and text—often lie near low-dimensional submanifolds due to physical constraints, as supported by both theoretical and empirical evidence. This motivates manifold learning, a key technique in machine learning and data analysis for handling high-

dimensional data. Despite its importance, privacy-preserving manifold learning remains largely unexplored. To address this gap, I aim to develop a differentially private manifold learning framework, building on tools from my previous research.

Bayesian Singular Learning Theory. Another exciting direction I plan to explore is Bayesian singular learning theory, which provides a rigorous statistical framework for analyzing complex models whose parameter spaces exhibit non-identifiability or singularities—properties often seen in deep neural networks. By leveraging tools from algebraic geometry, this framework captures how the shape and curvature of the parameter space influence generalization behaviour, even in the overparameterized regime. I am particularly interested in extending this theory to settings with intrinsic geometric structure, such as Riemannian or stratified manifolds, and integrating differential privacy constraints. This direction builds on my broader goal of understanding learning from both a statistical and geometric lens, and aligns naturally with recent trends in the theory of deep learning.

Incorporate TDA into Geometric Statistics/Machine Learning. Recent advances in topological data analysis (TDA) offer powerful tools for understanding the structure and generalization properties of modern learning systems. Notably, the work of Birdal [21] demonstrates that the persistent homology dimension of a network's training trajectory correlates strongly with its generalization error, providing a predictive and theoretically grounded metric for model performance. This suggests that persistent homology can serve not only as a descriptive invariant but also as a basis for topological regularization during training. Such approaches would complement existing geometric methods by addressing not only local curvature or metric properties of data manifolds but also their global topological invariants. Integrating TDA with differential geometry could further advance privacy-preserving mechanisms; for example, topological regularization might enhance the utility-privacy trade-off by preserving essential data topology while obscuring sensitive features. Building on my work in differential privacy on Riemannian manifolds, I aim to explore how TDA can inform the development of mechanisms that respect both geometric and topological constraints, particularly in applications like medical imaging, where data topology (e.g., connectivity, loops, voids) often encodes critical information. This direction would also extend naturally to Bayesian frameworks, where priors could incorporate topological features derived from persistent homology, or to manifold learning, where TDA tools might help identify latent topological structure in unlabeled data. By bridging TDA and geometric statistics, this research could yield interpretable, robust, and privacy-aware analytical tools for complex nonlinear data.

Develop Post-Selection Inference on General Manifold. Classical statistical methods assume that the research question is set before analyzing data. In practice, however, questions often emerge from the data itself, creating dependencies that challenge traditional inference guarantees. While this data-driven approach was once viewed skeptically, it is now widely accepted, particularly in model selection, where inference naturally depends on the chosen model. Post-selection inference addresses this challenge, but existing methods either apply only to specific selection procedures or sacrifice statistical power for generality. It has been demonstrated in [22] that building on concepts from the field of differential privacy, it's possible to derive selective confidence intervals that are both tractable computationally and powerful statistically. Motivated by this observation and built on our previous work dealing with differential privacy on general manifolds, I aim to extend post-selection inference to manifold-valued data.

References

[1] Xavier Pennec, Stefan Sommer, and Tom Fletcher. *Riemannian geometric statistics in medical image analysis*. Academic Press, 2019.

[2] Ian L. Dryden. Statistical analysis on high-dimensional spheres and shape spaces. *The Annals of Statistics*, 33(4):1643--1665, 2005.

[3] Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102--1123, 2009.

[4] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to bci classification. *Latent Variable Analysis and Signal Separation*, 09 2010.

[5] Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107--1116, 2018.

[6] Pavan K. Turaga and Anuj Srivastava. *Riemannian Computing in Computer Vision*. Springer Publishing Company, Incorporated, 1st edition, 2015.

[7] Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1--8, 2008.

[8] Guang Cheng and Baba C. Vemuri. A novel dynamic system in the space of spd matrices with applications to appearance tracking. *SIAM J. Img. Sci.*, 6(1):592–615, jan 2013.

[9] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399--2434, 2006.

[10] Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(37):1229--1250, 2013.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265--284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[12] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 94--103, 11 2007.

[13] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '07, page 273–282, New York, NY, USA, 2007. Association for Computing Machinery.

[14] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375--389, 2010.

[15] Matthew Reimherr and Jordan Awan. Kng: The k-norm gradient mechanism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] Matthew Reimherr, Karthik Bharath, and Carlos Soto. Differential privacy over riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 34, pages 12292--12303. Curran Associates, Inc., 2021.

[17] **Yangdi Jiang**, Lucy Mosquera, Bei Jiang, Linglong Kong, and Khaled El Emam. Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLOS ONE*, 17(6):1--19, 06 2022.

[18] **Yangdi Jiang**, Yi Liu, Xiaodong Yan, Anne-Sophie Charest, Linglong Kong, and Bei Jiang. Analysis of differentially private synthetic data: A measurement error approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21206--21213, Mar. 2024.

[19] **Yangdi Jiang**, Xiaotian Chang, Yi Liu, Lei Ding, Linglong Kong, and Bei Jiang. Gaussian differential privacy on riemannian manifolds. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[20] **Yangdi Jiang**, Xiaotian Chang, Lei Ding, Linglong Kong, and Bei Jiang. Exponential wrapped mechanism: A general and computationally efficient approach to differential privacy on hadamard manifolds. In *International Conference on Machine Learning*, In Review.

[21] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776--6789, 2021.

[22] Tijana Zrnic and Michael I. Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666 -- 1691, 2023.