RESEARCH STATEMENT: Hanbaek Lyu

My research focuses on discrete probability and nonconvex optimization, with particular interest in matrix factorization. The overarching goal of my work is to develop mathematical frameworks and practical tools to analyze large, complex systems composed of interconnected variables. I have been honored with two NSF grants supporting my work, which has been published in top journals and conferences such as *Nature Communications*, the *Annals of Probability*, *Forum of Mathematics Sigma*, *ICML*, and *NeurIPS*. I have delivered over 65 invited talks at prestigious institutions and conferences. I contribute to my research community by serving as a member of the editorial board for *Scientific Reports* and by organizing the 44th Midwest Probability Colloquium, as well as a minisymposium on "Probabilistic Methods in Machine Learning and Complex Systems" at SIAM MDS 2024. Furthermore, I have been devoted to promoting research experiences for undergraduates (REU) by designing and leading four NSF-funded REU programs, and mentoring 22 undergraduates, resulting in a publication in *Scientific Reports* and two manuscripts under review.

More specifically, by leveraging techniques from probability theory, machine learning, and optimization, I aim to address fundamental questions on large complex systems from three complementary perspectives: (1) model-driven, (2) data-driven, and (3) entropy-driven.

First, in an *interacting particle system*, particles interact on networks following simple rules that **model micro-scale mechanism of a specific complex system**. In order to understand the system's collective behavior, one needs to develop a thorough understanding of the tight inerplay between the network topology and the local interaction rules. I have investigated these systems across different contexts, such as coupled oscillators [21, 27, 22], chemical and neural media [9, 2], inter-species annihilation [6, 11, 7, 5, 1, 3, 10], and solitary waves in integrable systems [18, 12, 13, 19]. My work has addressed several longstanding questions in the field, offering new insights into these complex behaviors. In particular, the historic Korteweg-de Vries (KdV) equation of 1895 describes particle-like water waves called "solitons" and has been the focus of extensive research in mathematical physics and integrable systems. A major open question in KdV is determining the asymptotic lengths of solitons emerging from KdV with random initial conditions. In [18, 19], I have addressed this question for discrete versions of KdV and uncovered an intricate phase transition of the soliton lengths in initial particle distribution. A key technical innovation was the construction of an embedded semimartingale reflecting Brownian motion that tracks the evolution of interacting solitons.

Second, I develop universal tools for analyzing complex systems by focusing on **how they appear as large datasets**, rather than on how they are generated by specific mechanisms. *Networks* are ubiquitous objects that describe the dependence structures between entities in complex systems. My research has developed innovative methods for decomposing networks into a small set of key connection patterns and reconstructing the networks from them. A culmination of this research was published in *Nature Communications* (featured in the Editor's Highlight), where a reviewer praised it as a "tour-de-force" in the field. This research is grounded in matrix factorization, a powerful machine-learning tool used to uncover hidden structures in large datasets. My contributions include advancing optimization algorithms that handle data with complex dependencies and pushing the boundaries of what can be achieved by incorporating machine learning in network analysis. These works are published in top machine learning venues such as *ICML* [4, 14, 17, 20, 29, 15], *NeurIPS* [16], and *JMLR* [28, 24, 23].

Third, statistical physicists use *constrained random matrices* to model complex interactions in heavy atomic nuclei. The guiding philosophy is that, among all possible ways that nuclei may interact, the universe selects the **maximum entropy** (i.e., most likely) interaction mechanism at random given physical constraints. In [25], I have developed a theory for constrained random matrices that explains the most likely structure given specific row and column sums. I showed that the random entries somehow automatically and collectively form a highly ordered structure to maximize entropy. More surprisingly, this "maximum entropy structure" is closely related to entropic optimal transport, a topic of great interest in mathematical physics and machine learning. This is an entirely new and unexpected connection between random matrix theory and optimal transport, which many experts in the field have appreciated in my recent lectures. For instance, it has already triggered the new discovery that Sinkhorn's algorithm in optimal transport can be generalized to the constrained random matrix setting. I have developed the core idea over several years, addressing a longstanding conjecture in combinatorics regarding the structural phase transition of random contingency tables [8, 26].

As I continue pushing the boundaries of complex systems across these three complementary perspectives, more surprising connections and synergistic discoveries are expected to emerge.

References

[1] Kimberly Affeld, Christian Dean, Matthew Junge, Hanbaek Lyu, Connor Panish, and Lily Reeves. Four-parameter coalescing ballistic annihilation. *Journal of Statistical Physics*, 191(7):89, 2024.

[2] Ander Aguirre, Hanbaek Lyu, and David Sivakoff. Phase transition in one-dimensional excitable media with variable interaction range. *arXiv preprint arXiv:1701.00319*, 2024.

[3] Sungwon Ahn, Matthew Junge, Hanbaek Lyu, Lily Reeves, Jacob Richey, and David Sivakoff. Diffusion-limited annihilating-coalescing systems. *Under review in Electronic Journal of Probability (arXiv preprint arXiv:2305.19333)*, 2023.

[4] Ahmet Alacaoglu and Hanbaek Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *International Conference on Machine Learning*, pages 458–489. PMLR, 2023.

[5] Luis Benitez, Matthew Junge, Hanbaek Lyu, Maximus Redman, and Lily Reeves. Three-velocity coalescing ballistic annihilation. *Electronic Journal of Probability*, 28:1–18, 2023.

[6] Michael Damron, Janko Gravner, Matthew Junge, Hanbaek Lyu, and David Sivakoff. Parking on transitive unimodular graphs. *The Annals of Applied Probability*, 29(4):2089–2113, 2019.

[7] Michael Damron, Hanbaek Lyu, and David Sivakoff. Stretched exponential decay for subcritical parking times on. *Random Structures & Algorithms*, 59(2):143–154, 2021.

[8] Samuel Dittmer, Hanbaek Lyu, and Igor Pak. Phase transition in random contingency tables with non-uniform margins. *Transactions of the American Mathematical Society*, 373(12):8313–8338, 2020.

[9] Janko Gravner, Hanbaek Lyu, and David Sivakoff. Limiting behavior of 3-color excitable media on arbitrary graphs. *The Annals of Applied Probability*, 28(6):3324 – 3357, 2018.

[10] Tobias Johnson, Matthew Junge, Hanbaek Lyu, and David Sivakoff. Particle density in diffusion-limited annihilating systems. *The Annals of Probability*, 51(6):2301–2344, 2023.

[11] Matthew Junge and Hanbaek Lyu. The phase structure of asymmetric ballistic annihilation. *The Annals of Applied Probability*, 32(5):3797–3816, 2022.

[12] Atsuo Kuniba and Hanbaek Lyu. Large deviations and one-sided scaling limit of randomized multicolor box-ball system. *Journal of Statistical Physics*, 178(1):38–74, 2020.

[13] Atsuo Kuniba, Hanbaek Lyu, and Masato Okado. Randomized box–ball systems, limit shape of rigged configurations and thermodynamic bethe ansatz. *Nuclear Physics B*, 937:240–271, 2018.

[14] Dohyun Kwon and Hanbaek Lyu. Complexity of block coordinate descent with proximal regularization and applications to wasserstein cp-dictionary learning. In *International Conference on Machine Learning*, pages 18114–18134. PMLR, 2023.

[15] Jeongyeol Kwon, Dohyun Kwon, and Hanbaek Lyu. On the complexity of first-order methods in stochastic bilevel optimization. *International Conference on Machine Learning*, 2024.

[16] Joowon Lee, Hanbaek Lyu, and Weixin Yao. Exponentially convergent algorithms for supervised matrix factorization. In *Advances in Neural Information Processing Systems*, volume 36, pages 76947–76959. Curran Associates, Inc., 2023.

[17] Joowon Lee, Hanbaek Lyu, and Weixin Yao. Supervised matrix factorization: Local landscape analysis and applications. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26752–26788. PMLR, 2024.

[18] Lionel Levine, Hanbaek Lyu, and John Pike. Double jump phase transition in a soliton cellular automaton. *International Mathematics Research Notices*, 2022(1):665–727, 2022.

[19] Joel Lewis, Hanbaek Lyu, Pavlo Pylyavskyy, and Arnab Sen. Scaling limit of soliton lengths in a multicolor box-ball system. *To appear in Form of Mathematics, Sigma*, 2024.

[20] Yuchen Li, Laura Balzano, Deanna Needell, and Hanbaek Lyu. Convergence and complexity guarantee for inexact first-order Riemannian optimization algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27376–27398. PMLR, 2024.

[21] Hanbaek Lyu. Synchronization of finite-state pulse-coupled oscillators. *Physica D: Nonlinear Phenomena*, 303:28–38, 2015.

[22] Hanbaek Lyu. Time complexity of synchronization of discrete pulse-coupled oscillators on trees. *arXiv preprint arXiv:1610.00837*, 2023.

[23] Hanbaek Lyu. Stochastic regularized majorization-minimization with weakly convex and multi-convex surrogates. *Journal of Machine Learning Research*, 25(306):1–83, 2024.

[24] Hanbaek Lyu, Facundo Memoli, and David Sivakoff. Sampling random graph homomorphisms and applications to network data analysis. *Journal of Machine Learning Research*, 24:1–79, 2023.

[25] Hanbaek Lyu and Sumit Mukherjee. Concentration and limit of large random matrices with given margins. *arXiv preprint arXiv:2407.14942*, 2024.

[26] Hanbaek Lyu and Igor Pak. On the number of contingency tables and the independence heuristic. *Bulletin of the London Mathematical Society*, 54(1):242–255, 2022.

[27] Hanbaek Lyu and David Sivakoff. Persistence of sums of correlated increments and clustering in cellular automata. *Stochastic Processes and their Applications*, 2018.

[28] Hanbaek Lyu, Christopher Strohmeier, and Deanna Needell. Online tensor factorization and cp-dictionary learning for markovian data. *Journal of Machine Learning Research 23 (2022) 1-50*, 2022.

[29] William Powell and Hanbaek Lyu. Stochastic optimization with arbitrary recurrent data sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41000–41038. PMLR, 2024.

HANBAEK LYU, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF WISCONSIN - MADISON
Email address: `hlyu@math.wisc.edu`