Yijun Dong

Curriculum Vitae

✉ *yd1319@nyu.edu*
dyjdongyijun.github.io

Research Interests

Randomized Numerical Linear Algebra, Learning Theory.
I am interested in the computational and sample efficiency of algorithms in machine learning and scientific computing. From the computational efficiency perspective, my work is centered on matrix sketching and randomized low-rank decompositions like SVD and CUR. From the sample efficiency perspective, my work focuses on the generalization and distributional robustness of learning algorithms in data-limited settings.

Employment

2023-Present **New York University, Courant Institute**, New York, NY, US,
Assistant Professor/Courant Instructor.

Education

2018-2023 **University of Texas at Austin, Oden Institute**, Austin, TX, US,
Ph.D. in Computational Science, Engineering, and Mathematics.
- Advisors: Prof. Per-Gunnar Martinsson, Prof. Rachel Ward.
- Thesis: Randomized Dimension Reduction with Statistical Guarantees.

2014-2018 **Emory University**, Atlanta, GA, US,
B.S. in Applied Mathematics & Engineering Science, Magna Cum Laude.
- Advisors: Prof. Effrosyni Seitaridou, Prof. Eric Weeks.
- Thesis: Crystals and Liquids in Gravitationally Confined Quasi-2D Colloidal Systems.

Publications (∗ denotes equal contribution or alphabetical order)

Conference Publications

- Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning. Yijun Dong*, Hoang Phan*, Xiang Pan*, Qi Lei. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. (to appear)
- Cluster-aware Semi-supervised Learning: Relational Knowledge Distillation Provably Learns Clustering. Yijun Dong*, Kevin Miller*, Qi Lei, Rachel Ward. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- Adaptively Weighted Data Augmentation Consistency Regularization for Robust Optimization under Concept Shift. Yijun Dong*, Yuege Xie*, Rachel Ward. *International Conference on Machine Learning (ICML)*, 2023.
- Sample Efficiency of Data Augmentation Consistency Regularization. Shuo Yang*, Yijun Dong*, Rachel Ward, Inderjit S Dhillon, Sujay Sanghavi, Qi Lei. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Journal Publications

- Efficient Bounds and Estimates for Canonical Angles in Randomized Subspace Approximations.

Yijun Dong, Per-Gunnar Martinsson, Yuji Nakatsukasa. *SIAM Journal on Matrix Analysis and Applications*, 2024.

- Simpler is better: A comparative study of randomized algorithms for computing the CUR decomposition. Yijun Dong, Per-Gunnar Martinsson. *Advances in Computational Mathematics*, 2023.
- Quantifying Biofilm Formation of *Sinorhizobium meliloti* Bacterial Strains in Microfluidic Platforms by Measuring the Diffusion Coefficient of Polystyrene Beads. Chen Cheng*, Yijun Dong*, Matthew Dorian*, Farhan Kamili*, Effrosyni Seitaridou. *Open Journal of Biophysics*, 2017.

Preprints & Workshop Papers

- Randomly Pivoted V-optimal Design: Fast Data Selection under Low Intrinsic Dimension. Yijun Dong*, Xiang Pan*, Hoang Phan*, Qi Lei. *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.
- Greedy Output Approximation: Towards Efficient Structured Pruning for LLMs Without Retraining. Jianwei Li, Yijun Dong, Qi Lei. *arXiv:* 2407.19126, 2024.
- Robust Blockwise Random Pivoting: Fast and Accurate Adaptive Interpolative Decomposition. Yijun Dong, Chao Chen, Per-Gunnar Martinsson, Katherine Pearce. *arXiv:* 2309.16002, 2023.
- Adaptive Parallelizable Algorithms for Interpolative Decompositions via Partially Pivoted LU. Katherine Pearce, Chao Chen, Yijun Dong, Per-Gunnar Martinsson. *arXiv:* 2310.09417, 2023.

Teaching Experience

2023-Present **Instructor**, Courant Institute, New York University, New York, NY.
- Fall 2024: Computational Statistics.
- Spring 2024: Mathematics for Economics.
- Fall 2023: Discrete Mathematics.

2023/07 **Teaching Assistant**, Simons Laufer Mathematical Sciences Institute (SLMath) Summer Graduate School, IBM Almaden, San Jose, CA.
- Mathematics of Big Data: Sketching and (Multi-) Linear Algebra (TA for Drs. Kenneth Clarkson, Lior Horesh, Misha Kilmer, Tamara Kolda, and Shashanka Ubaru).

2020-2022 **Teaching Assistant**, UT Austin, Austin, TX.
- Fall 2022: Differential Equations with Linear Algebra (TA for Dr. Michael Novack).
- Fall 2021: Numerical Analysis: Linear Algebra (TA for Prof. Per-Gunnar Martinsson).
- Fall 2020: Differential Equations with Linear Algebra (TA for Prof. Sam Raskin).

2015-2016 **Student Tutor**, Emory University, Oxford, GA.
- Introduction to Physics, Modern Physics.

Service

Review
- Journals: Annals of Applied Probability, BIT Numerical Mathematics, Calcolo, IEEE Transactions on Signal Processing, IMA Journal of Numerical Analysis, Journal of Computational Mathematics and Data Science, SIAM Journal on Matrix Analysis and Applications, SIAM Journal on Scientific Computing.
- Conferences: AISTATS (2023), NeurIPS (2024).

Organization
- SIAM MDS24 minisymposium on "Efficient Computation and Learning with Randomized Sampling and Pruning", with Yifan Chen, Qi Lei.

Talks

- Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning
  - 2025/01 Joint Mathematics Meetings (JMM 2025) ILAS Special Session on Randomness in Numerical Linear Algebra, Seattle, Washington.
  - 2024/12 Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada. (poster)
  - 2024/10 University of Delaware Numerical Analysis and PDE seminar, Newark, Delaware. (invited)
- RBRP: Fast and Accurate Interpolation Decomposition with Adaptiveness and Randomness
  - 2024/10 SIAM Conference on Mathematics of Data Science (MDS24) minisymposium on "Efficient Computation and Learning with Randomized Sampling and Pruning", Atlanta, Georgia. (poster)
  - 2024/03 SIAM Conference on Parallel Processing for Scientific Computing (PP24) minisymposium on "Randomized Methods in Linear Solvers and Matrix Factorizations", Baltimore, Maryland.
- Relational Knowledge Distillation Provably Learns Clustering
  - 2023/12 Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, Louisiana. (poster)
- Efficient Bounds and Estimates for Canonical Angles in Randomized Subspace Approximations
  - 2023/11 6th SIAM Texas-Louisiana Sectional Meeting (SIAM TX-LA 2023) minisymposium on "Nonlinear Algebra in Applications", Lafayette, Louisiana.
  - 2023/08 The International Council for Industrial and Applied Mathematics (ICIAM) minisymposium on "Randomized Numerical Linear Algebra", Tokyo, Japan.
  - 2023/03 Texas Women in Math Symposium (TWIMS2023), Austin, Texas.
- Adaptively Weighted Data Augmentation Consistency Regularization for Robust Optimization under Concept Shift
  - 2023/07 International Conference on Machine Learning (ICML 2023), Honolulu, Hawaii. (poster)
  - 2023/04 2023 Rising Stars in Computational and Data Sciences, Austin, Texas.
  - 2022/11 IPAM Workshop IV: Multi-Modal Imaging with Deep Learning and Modeling (CM-SWS4), Los Angeles, California. (poster)
- Sample Efficiency of Data Augmentation Consistency Regularization
  - 2023/04 IFML Workshop 2023, Seattle, Washington.
  - 2022/10 Oden Institute CSEM Student Forum, Austin, Texas.
  - 2022/09 SIAM Conference on Mathematics of Data Science (MDS22), San Diego, California. (poster)
- A comparative study of randomized algorithms for computing the CUR decomposition
  - 2022/01 Jane Street Symposium 2022, Virtual.
  - 2021/05 SIAM Conference on Applied Linear Algebra (LA21), Virtual. (poster)
- Forming 2D colloidal crystals with sedimented colloids
  - 2018/03 American Physical Society March Meeting, Los Angeles, California.

Fellowships and Awards

2024 SIAM PP24 Early Career Travel Award
2023 Graduate School Summer Fellowship, UT Austin
2023 Rising Stars in Computational and Data Sciences
2019 NIMS Graduate Fellowship, UT Austin
2018 Peter O'Donnell Graduate Fellowship, UT Austin
2018 Trevor Evans Award, Emory University
2017 Phi Beta Kappa & Sigma Pi Sigma, Emory University

Industrial Experience

2022/06-2022/08 **Dell Technologies**, Research Intern, Austin, TX.
- Semi-supervised tabular learning with data augmentation and consistency regularization

2021/05-2021/08 **Dell Technologies**, Research Intern, Austin, TX.
- Streaming telemetry time series compression on edge devices

Skills

Programming
- Proficient: Bash, Git, LaTeX, MATLAB, Python.
- Prior knowledge: C++, IDL, Java, Julia.

Language
- Chinese (native), English (proficient), Japanese