Research Statement

Yijun Dong (NYU Courant)

I work at the intersection of applied mathematics and machine learning (ML) theory. Leveraging tools from randomized algorithms, high-dimensional probability, and statistical learning theory, my research seeks to *develop a unified theoretical and algorithmic framework for computation- and sample-efficient methods in high-dimensional learning problems*.

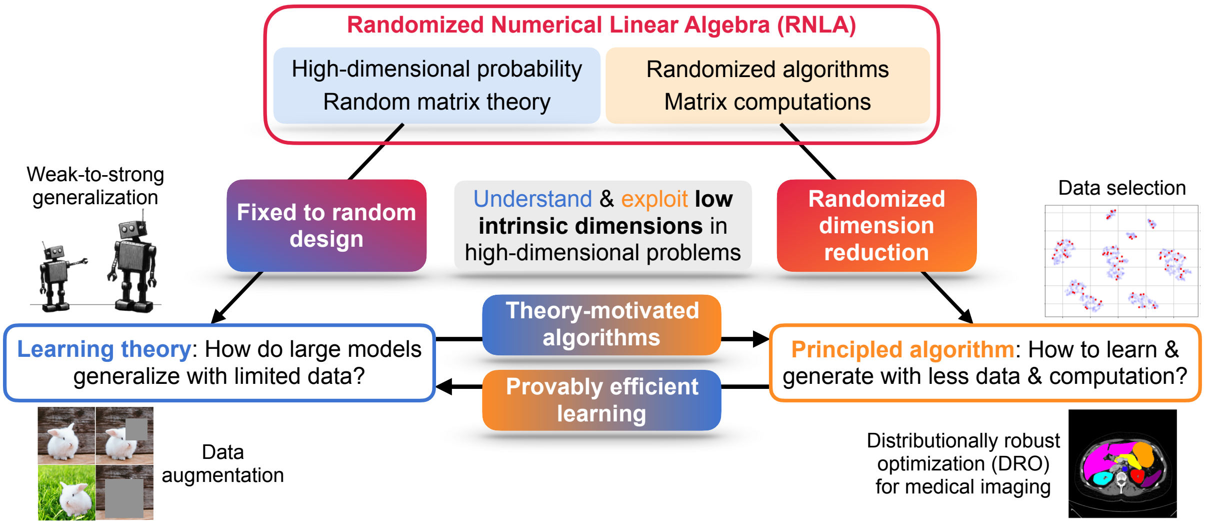

Randomized Numerical Linear Algebra (RNLA)
High-dimensional probability
Random matrix theory
Randomized algorithms
Matrix computations
Weak-to-strong generalization
Fixed to random design
Understand & exploit low intrinsic dimensions in high-dimensional problems
Randomized dimension reduction
Data selection
Learning theory: How do large models generalize with limited data?
Theory-motivated algorithms
Provably efficient learning
Principled algorithm: How to learn & generate with less data & computation?
Data augmentation
Distributionally robust optimization (DRO) for medical imaging

Figure 1: Three synergistic pillars of my research: **randomized numerical linear algebra** (Section 1), **learning theory** (Section 2), and **principled algorithms** (Section 3).

A central theme of my work is a modern manifestation of Occam's razor: conditioned on informative priors like large data matrices and foundation models, the simplest hypothesis is usually the best. *Low intrinsic dimensions* of high-dimensional problems brought by such virtue of simplicity pave the way toward understanding and designing learning algorithms. Building on this central thread, Figure 1 outlines three synergistic pillars of my research:

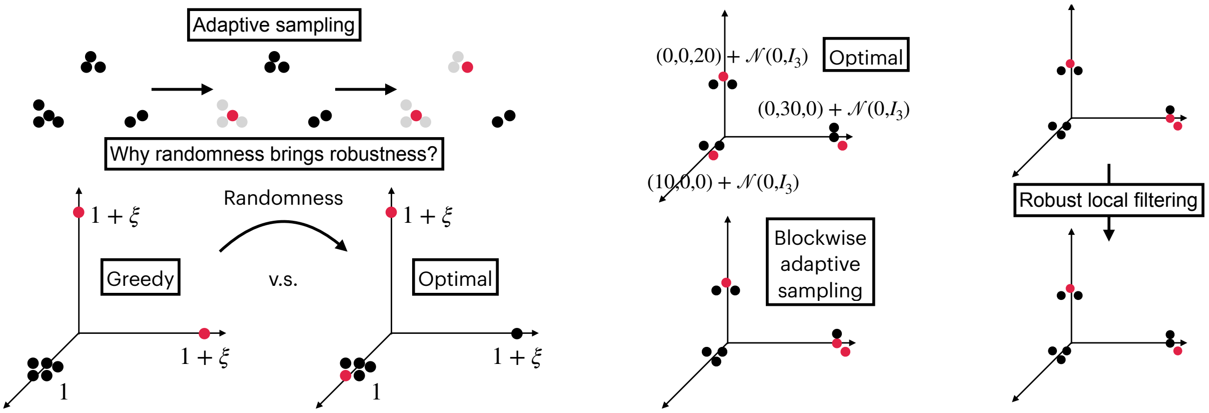
- **Randomized numerical linear algebra (RNLA)** (Section 1): I design fast and robust randomized algorithms for matrix low-rank approximations, including the *CUR/interpolative decomposition (ID)* [2, 8, 13] and *singular value decomposition (SVD)* [1, 5]. The mathematical and algorithmic foundations of RNLA have been instrumental in my work on understanding and designing learning algorithms (*e.g.*, [7, 9]) in Sections 2 and 3.
- **Learning theory** (Section 2): I develop theoretical underpinnings for when and why large models generalize with limited data in various learning paradigms like *data augmentation* [16], *knowledge distillation* [3], *weak-to-strong generalization* [9, 12], and *chain-of-thought* [15]. By unveiling the mechanisms behind these sample-efficient learning paradigms, my research in learning theory casts light on the algorithmic designs in Section 3 (*e.g.*, [4, 6, 7]).
- **Principled algorithms** (Section 3): I design robust and provable algorithms for computation- and sample-efficient learning, *e.g.*, *data selection* [6, 7, 14], *distributionally robust optimization* with applications to medical imaging [4], *structured pruning* of large language models (LLMs) [11], and *randomized time integration* for differential equations [10].

Sections 1 to 3 will focus on one representative topic in each pillar.

1 Randomized Numerical Linear Algebra (RNLA)

Data-driven algorithms are typically bottlenecked by the prohibitive costs of processing and storing large-scale data matrices. Real-world high-dimensional problems frequently exhibit latent low-dimensional structures [38], inducing different notions of *low intrinsic dimensions* [18, 34, 36] and motivating dimensionality reduction methods like low-rank approximations. My research in RNLA leverages two core randomization techniques—*sampling* [8] and *sketching* [2, 5, 13]—to exploit such low intrinsic dimensions and design fast, accurate, and robust algorithms for low-rank approximations.

1.1 Robust Blockwise Adaptive Sampling for Interpolative Decomposition (ID). *ID* is a low-rank approximation that selects a subset of rows/columns as a basis. It has wide applications in numerical analysis and ML, including rank-structured matrix compression [29], low-rank adaptation (LoRA) [30], and data selection [6, 7]. Although polynomial-time heuristics with near-optimal accuracy (*e.g.*, [22, 26]) exist for the NP-hard optimal subset selection problem [24], the unprecedented scales of modern computational challenges call for efficiency beyond asymptotic complexities, *e.g.*, *hardware efficiency* on parallel processors and *rank-adaptiveness* supporting adaptive selection with automatic termination.


Adaptive sampling
Why randomness brings robustness?
$1+\xi$
Randomness
Greedy
v.s.
Optimal
$1+\xi$
1
$1+\xi$
$1+\xi$
1
$(0,0,20)+\mathcal{N}(0,I_3)$
Optimal
$(0,30,0)+\mathcal{N}(0,I_3)$
$(10,0,0)+\mathcal{N}(0,I_3)$
Blockwise adaptive sampling
Robust local filtering

(a) Adaptive sampling and its robustness (b) Suboptimality of blockwise adaptive sampling (c) Fix with RBRP

Figure 2: Illustrations of RBRP for row (column) subset selection. Dots denote n rows in $\mathbb{R}^{n\times d}$; red dots are rows selected by ID; and gray dots are rows removed by adaptive updates.

With Professor Gunnar Martinsson's research group at UT Austin, I developed *robust blockwise random pivoting (RBRP)* [8], a randomized ID algorithm that is hardware-efficient, rank-adaptive, and robust to adversarial inputs, with state-of-the-art accuracy and efficiency. Figure 2 illustrates the key ideas of RBRP: (a) Adaptive sampling [22] achieves near-optimal accuracy by combining adaptiveness and randomness, circumventing adversarial inputs like Figure 2(a). (b) However, the sequential nature of adaptive sampling compromises hardware efficiency, while blockwise adaptive sampling is vulnerable to inputs like Figure 2(b). (c) RBRP resolves this tension through *robust local filtering* that removes redundant points in each block with negligible additional cost (Figure 2(c)), achieving hardware efficiency, rank-adaptiveness, and robustness.

1.2 Sketching-based Low-rank Approximations with Posterior Guarantees. Alternative to sampling, *sketching* [28, 40] via Johnson-Lindenstrauss transforms [32] randomly projects high-dimensional data to a low-dimensional subspace while preserving essential information. Sketching has been thoroughly studied for low-rank approximations with a priori guarantees [28, 40]. However, data-agnostic a priori guarantees from concentration bounds are often too pessimistic for practical error estimation, while data-dependent *posterior* guarantees are much less explored. My Ph.D. work [1, 2, 5] with Professors Gunnar Martinsson and Yuji Nakatsukasa (Oxford) developed and analyzed sketching-based randomized algorithms for ID, CUR, and SVD with tight and efficiently computable posterior error guarantees.

2 Learning Theory

A central question of ML theory is *when and why large models generalize in the overparametrized regime* (*i.e.*, with fewer samples $n < d$ than the parameter count d). Benign overfitting [37] provides a canonical insight: when the essential features of data concentrate in a subspace of low dimension $k_* \ll n$, overparametrized models can fit the n samples while generalizing well by diluting noise in the $(d - k_*)$-dimensional subspace orthogonal to essential features. My research in learning theory explains the mechanisms behind various sample-efficient learning paradigms whose low intrinsic dimensions k_* arise implicitly from the simplicity bias [9, 12] or explicitly from structure-aware objectives [3, 16].

2.1 Intrinsic Dimensions of Fine-tuning Explain Weak-to-Strong (W2S) Generalization. Post-training alignment adapts pre-trained models to downstream tasks using limited data, commonly through highly overparametrized *fine-tuning*. The sample efficiency of fine-tuning is empirically justified by its low intrinsic dimensions [17]. Conditioned on the simple dynamics of fine-tuning in the kernel regime [31, 35], low intrinsic dimensions emerge from the simplicity bias of optimizers like gradient descent [41]. Low intrinsic dimensions provide fresh perspectives for assorted post-training schemes, like W2S generalization [20].

As contemporary ML models outperform human in many domains, an imminent question of *superalignment* asks whether stronger models can still learn from weaker human supervision. W2S generalization [20] gives a positive answer: a strong pre-trained student fine-tuned with pseudo-labels generated by a weaker teacher often outperforms its teacher. Collaborating with undergraduates Yicheng Li (NYU Shanghai), Yunai Li (SJTU), Professors Jason D. Lee (UC Berkeley), and Qi Lei (NYU), I developed a theoretical framework to explain W2S from a variance reduction perspective, via the low intrinsic dimension of fine-tuning [9]. We model fine-tuning the strong student f_s and weak teacher f_w as regression problems over high-dimensional pre-trained features concentrated in low-dimensional subspaces, $\mathcal{V}_s, \mathcal{V}_w$ of dimensions d_s, d_w, respectively. Assuming that the pre-trained features of both teacher and student are sufficiently expressive for the downstream task (*i.e.*, negligible bias), we characterize the dominant variance of W2S precisely: when fine-tuning the strong student with $N \gtrsim d_s$ pseudo-labels generated by the weak teacher fine-tuned with $n \gtrsim d_w$ noisy labels with variance σ^2,

$$\mathbf{Var}(f_{\mathrm{w2s}}) \asymp \sigma^2 \Bigg(\underset{\text{Var. in } \mathcal{V}_s \cap \mathcal{V}_w}{\boxed{\frac{d_{s\wedge w}}{n}}} + \underset{\textbf{W2S}}{\boxed{\frac{d_s}{N}}} \underset{\text{Var. in } \mathcal{V}_w \setminus \mathcal{V}_s}{\boxed{\frac{d_w - d_{s\wedge w}}{n}}} \Bigg), \tag{1}$$

where $d_{s\wedge w} \in [0, \min\{d_s, d_w\}]$ measures $\mathcal{V}_s, \mathcal{V}_w$ overlap through their geodesic distance. (1) implies that variance of the teacher is inherited by the student in $\mathcal{V}_s \cap \mathcal{V}_w$, while reduced by a factor of d_s/N in the subspace of discrepancy $\mathcal{V}_w \setminus \mathcal{V}_s$. W2S generalization emerges from such variance reduction when the student and teacher are sufficiently different (*i.e.*, small $d_{s\wedge w}$).

2.2 Structure-awareness Unveils Intrinsic Dimensions. Besides simplicity biases, structure-aware objectives can explicitly induce low intrinsic dimensions and enable sample-efficient learning. With Shuo Yang, Kevin Miller, Professors Rachel Ward, Inderjit Dhillon, Sujay Sanghavi (UT Austin), and Qi Lei, I provided theoretical underpinnings for the sample efficiency of two such structure-aware learning schemes: *data augmentation consistency regularization* [16] and *relational knowledge distillation* [3].

3 Principled Learning Algorithms

The algorithmic and theoretical foundations in Sections 1 and 2 have been instrumental in my work on efficient, robust, and provable algorithms for large-scale learning problems.

3.1 Efficient Data Selection for Fine-tuning. The theoretical insights in Section 2.1 suggest that sample efficiency of fine-tuning stems from its low intrinsic dimensions. A pre-trained foundation model is an "Occam's razor" that concentrates high-dimensional downstream features in low-dimensional subspaces. However, downstream data are often redundant, while labels tend to be noisy and expensive. This calls for efficient strategies to find high-quality data subsets, labeling and fine-tuning on which preserve the generalization of full-data training.

With Professor Qi Lei's research group at NYU, I developed a scalable data selection framework, *Sketchy Moment Matching (SkMM)* [7], that achieves near-optimal sample complexities provably and state-of-the-art performance empirically. SkMM controls both bias and variance induced by data selection in two stages: (i) Leveraging algorithmic tools from RNLA, bias is controlled via *gradient sketching* that explores the fine-tuning parameter space of high dimension d for an informative subspace $\mathcal{S}$ of intrinsic dimension $\dim(\mathcal{S}) \ll d$. (ii) Variance is reduced over the low-dimensional $\mathcal{S}$ via *moment matching* that accelerates classical optimal experimental design methods through continuous relaxations [7] or RNLA-inspired heuristics based on adaptive sampling [6]. Theoretically, fine-tuning on n samples selected by SkMM preserves the fast-rate generalization $O(\dim(\mathcal{S})/n)$, with sample complexities independent of d. Algorithmically, SkMM runs in linear time $O(Nd)$ for a full data size N. The computational bottleneck is a single epoch of backpropagation for gradient sketching that is parallelizable on modern hardware and can be further accelerated with structured sketching (*e.g.*, [23]). Empirically, SkMM demonstrates state-of-the-art computation- and sample-efficiency against a broad spectrum of baselines on assorted regression and classification tasks.

3.2 Applications in Scientific ML and Beyond. Benefits of randomization extend beyond efficiency to other desiderata like *stability*. Collaborating with Professor Benjamin Peherstorfer's research group at NYU, I developed a randomized time integration scheme for solving evolution problems like dynamical systems and partial differential equations (PDEs) using nonlinear parametrizations like neural networks [10]. Our randomized time integrator uses sketching as an efficient regularization that provably stabilizes the poorly conditioned least-squares problems that arise from nonlinear parametrization. In addition, my learning theory work in Section 2 have inspired practical *distributionally robust algorithms*. For example, my Ph.D. work [4] with Professor Rachel Ward's research group at UT Austin introduced a scalable distributionally robust optimization algorithm for reliable medical image segmentation under concept shifts.

4 Ongoing and Future Research Agenda

An overarching goal of my research is to build mathematical underpinnings and algorithmic improvements for high-dimensional learning problems by bridging them with theoretical and algorithmic wisdom in RNLA. My future research will be built around the three synergistic pillars in Figure 1, extending foundations and insights in Sections 1 to 3 to emerging challenges in large-scale learning[1].

4.1 RNLA in Large-Scale Optimization and Learning. Large-scale training of ML models involves two core questions: (i) how to optimize the model with low per-iteration costs and fast convergence, and (ii) whether the algorithm steers the model toward global minima that

[1] Sections 4.2 and 4.3 are the themes of my MSCA-PF proposal, "SIMPLE-LLMs: Simplicity-Inspired Mechanistic Principles for Learning and Emergence in LLMs", hosted by Professor Nicolas Boullé at Imperial College.

generalize well? Toward the first question, low/mixed-precision arithmetic is becoming a de facto choice for reducing per-iteration costs in large-scale training [25]. This reinvigorates a classical theme in numerical linear algebra: *numerical stability*. Motivated by our finding in [10] that randomization stabilizes poorly conditioned problems efficiently, an exciting future direction is using randomization to design fast and stable matrix-preconditioned stochastic optimizers, like variants of Shampoo [27] and Muon [33], for low-precision training. A parallel but equally important direction posed by the second question is to understand the learning dynamics of different preconditioned optimizers. Bridging my experience in RNLA and learning theory, I plan to explore these two directions jointly, pursuing fast and stable preconditioned optimizers with favorable generalization.

4.2 Sample Efficiency of In-context Learning (ICL). Alternative to fine-tuning, ICL has emerged as a powerful post-training alignment scheme that adapts pre-trained LLMs to downstream tasks with only a few in-context demonstrations during inference, without updating model parameters [19]. A canonical insight exploited in Sections 2.1 and 3.1 is that sample efficiency of post-training alignment stems from the "Occam's razor" brought by rich knowledge encoded in LLMs. This motivates an exciting question: *what is the "Occam's razor" of ICL, and how does it emerge from LLMs?* I envision quantifying the sample efficiency of ICL through intrinsic low-complexity structures unveiled by LLMs, and designing concise and reliable contexts for ICL leveraging such mechanisms.

4.3 Efficient Matrix Computations during Post-training. Besides sample efficiency, "Occam's razor" in post-training further facilitate progress toward *computational and memory efficiency*. (i) For fine-tuning, the presence of low intrinsic dimensions explains the empirical success of *low-rank adaptation (LoRA)* [30], which explicitly enforces "simplicity" of fine-tuning updates by constraining them to low-rank matrices. This opens up exciting opportunities to bridge the understanding of training dynamics from learning theory with algorithmic building blocks from RNLA, exploring alternative low-complexity structures for LoRA that steer fine-tuning toward better generalization and efficiency. (ii) For inference of LLMs, the attention mechanism is a computational bottleneck that scales quadratically with the context length [39]. Motivated by common low-complexity patterns in attention matrices [42], sparse/low-rank attention approximations have been exploited to alleviate the computational and memory burden [21]. My ongoing collaboration with a group of applied mathematicians led by Professors Laura Grigori (EPFL), Anna Ma (UCI), and Deanna Needell (UCLA), initiated during the IPAM RNLA Workshop at UCLA, pursues lightweight LLM inference leveraging fast matrix computations from RNLA.

References I: My Publications

[1] Yijun Dong. Randomized dimension reduction with statistical guarantees. *Ph.D. Thesis, University of Texas at Austin*, 2023.

[2] Yijun Dong and Per-Gunnar Martinsson. Simpler is better: a comparative study of randomized pivoting algorithms for cur and interpolative decompositions. *Advances in Computational Mathematics*, 49(4):66, 2023.

[3] Yijun Dong, Kevin Miller, Qi Lei, and Rachel Ward. Cluster-aware semi-supervised

learning: Relational knowledge distillation provably learns clustering. *Advances in Neural Information Processing Systems*, 36:40799–40831, 2023.

[4] Yijun Dong, Yuege Xie, and Rachel Ward. Adaptively weighted data augmentation consistency regularization for robust optimization under concept shift. In *International Conference on Machine Learning*, pages 8296–8316, 2023.

[5] Yijun Dong, Per-Gunnar Martinsson, and Yuji Nakatsukasa. Efficient bounds and estimates for canonical angles in randomized subspace approximations. *SIAM Journal on Matrix Analysis and Applications*, 45(4):1978–2006, 2024.

[6] Yijun Dong, Xiang Pan, Hoang Phan, and Qi Lei. Randomly pivoted v-optimal design: Fast data selection under low intrinsic dimension. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

[7] Yijun Dong, Viet Hoang Phan, Xiang Pan, and Qi Lei. Sketchy moment matching: Toward fast and provable data selection for finetuning. *Advances in Neural Information Processing Systems*, 37:43367–43402, 2024.

[8] Yijun Dong, Chao Chen, Per-Gunnar Martinsson, and Katherine Pearce. Robust blockwise random pivoting: Fast and accurate adaptive interpolative decomposition. *SIAM Journal on Matrix Analysis and Applications*, 46(3):1791–1815, 2025.

[9] Yijun Dong, Yicheng Li, Yunai Li, Jason D Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. *International Conference on Machine Learning*, 2025.

[10] Yijun Dong, Paul Schwerdtner, and Benjamin Peherstorfer. Randomized time stepping of nonlinearly parametrized solutions of evolution problems. *Manuscript in preparation*, 2025.

[11] Jianwei Li, Yijun Dong, and Qi Lei. Greedy output approximation: Towards efficient structured pruning for llms without retraining. *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025.

[12] Chenruo Liu, Yijun Dong, and Qi Lei. Does Weak-to-strong Generalization Happen under Spurious Correlations? *Manuscript in preparation*, 2025.

[13] Katherine Pearce, Chao Chen, Yijun Dong, and Per-Gunnar Martinsson. Adaptive parallelizable algorithms for interpolative decompositions via partially pivoted lu. *Numerical Linear Algebra with Applications*, 32(1):e70002, 2025.

[14] Viet Hoang Phan, Yijun Dong, Andrew Gordon Wilson, and Qi Lei. Balanced locality-sensitive hashing for online data selection. *Manuscript in preparation*, 2025.

[15] Zihan Wang, Yijun Dong, and Qi Lei. When does Chain-of-Thought Help: A Markovian Perspective. *Manuscript in preparation*, 2025.

[16] Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 3825–3853. PMLR, 2023.

References II: Related Works

[17] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, 2021.

[18] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning*, pages 4971–5012. PMLR, 2024.

[21] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. *Advances in Neural Information Processing Systems*, 34:17413–17426, 2021.

[22] Yifan Chen, Ethan N Epperly, Joel A Tropp, and Robert J Webber. Randomly pivoted cholesky: Practical approximation of a kernel matrix with few entry evaluations. *Communications on Pure and Applied Mathematics*, 78(5):995–1041, 2025.

[23] Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong. Optimal subspace embeddings: Resolving nelson-nguyen conjecture up to sub-polylogarithmic factors. *arXiv preprint arXiv:2508.14234*, 2025.

[24] Ali Civril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

[25] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

[26] Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[27] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[28] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[29] Kenneth L Ho and Leslie Greengard. A fast direct solver for structured linear systems by recursive skeletonization. *SIAM Journal on Scientific Computing*, 34(5):A2507–A2532, 2012.

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[31] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[32] William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[33] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[35] Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pages 23610–23641. PMLR, 2023.

[36] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[37] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[38] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[41] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[42] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.