

Scalably computing metric magnitude

Steve Huntsman

STEVE.HUNTSMAN@CYNNOVATIVE.COM

Jewell Thomas

JEWELL.THOMAS@ALUMNI.UNC.EDU

Cynthia Ukawu

CYNTHIA.UKAWU@CYNNOVATIVE.COM

Abstract

Applications of metric magnitude often rely on numerically exact results in order to exploit a connection with information theory. We examine various approaches for scaling the dense linear algebra involved and identify hierarchical low-rank solvers as a preferred approach, with a clear path to scales of 10^5 points on a single powerful workstation, and larger scales using our containerized CUDA-enabled C++/MPI pipeline.

Keywords: Magnitude, structured kernel matrix, STRUMPACK

1. Introduction

The theory of magnitude (see §2 for a quick overview) generalizes notions of size and Euler characteristic to the context of enriched categories (Leinster, 2013; Leinster and Meckes, 2016; Leinster, 2021). It has already seen remarkable though still preliminary applications in the context of metric spaces, especially of strict negative type, where it generalizes classical information-theoretical constructions to incorporate notions of geometry.

However, applications of metric magnitude have often not scaled, and the techniques have not found wider adoption, primarily because they require solving dense linear systems of the form $Zw = 1$ with $Z_{jk} = \exp(-td_{jk})$ and d a distance matrix. Yet the *weighting vectors* w are exceptionally versatile, with uses for virtually any task involving analysis of Euclidean point clouds and/or information-theoretical characterizations. Accordingly, Andreeva et al. (2025) introduced both iterative and greedy schemes for approximating weighting vectors, but neither appears suitable for applications that substantially invoke information-theoretical properties. The iterative scheme produces a nonnegative weighting approximation that is bound to have components with the wrong sign near the most natural and important regime, and the greedy scheme ignores many points without guarantees.

We have therefore undertaken to directly address scalability bottlenecks for numerically exact solutions, in line with work on kernel ridge regression (Chávez et al., 2020). We consider three approaches: sparsification, hierarchical low-rank solvers for structured kernel matrices (Hackbusch, 2015), and (detailed in §B) Nyström approximation. We have determined that while these approaches cannot be fruitfully exploited in conjunction and two of them fail badly, structured kernel matrix solvers show real promise.

There is an easy bound $t_+ \leq \log(n-1)/\min_j \min_{k \neq j} d_{jk}$ for the interval $[t_+, \infty)$ where a natural information-theoretical interpretation of w is available (Huntsman, 2023a), but determining t_+ precisely requires solving $Zw = 1$ for different values of t . In this paper, we

focus on this problem (and retaining w as a byproduct) for data in $\mathbb{R}^m$. Figure 1 shows w for both the bound above and for t_+ itself on the same underlying points.¹

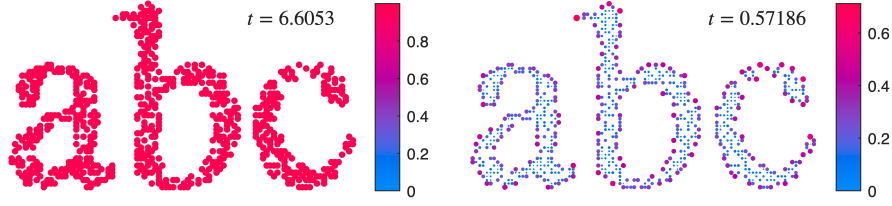


Figure 1: Left: w for $t = \log(n - 1) / \min_j \min_{k \neq j} d_{jk} \approx 6.605$, where d is the Euclidean distance on $n = 740$ unique points in $\mathbb{R}^2$ obtained by sampling 1000 points uniformly with replacement from a complicated support over $\mathbb{Z}^2$. Components of w are indicated by size and color: note that all are nearly 1. Right: as in the left panel, with the same points and distance matrix, but for $t = t_+ \approx 0.572$.

The upside of scalably solving $Zw = 1$ is highlighted by demonstrations and applications to, e.g., classification, active learning, and anomaly detection (Bunch et al., 2020); computer vision (Adamer et al., 2024); sparse reinforcement learning (Guo et al., 2021); introspection of neural representations (Bunch et al., 2021; Andreeva et al., 2023; Limbeck et al., 2024); pooling in graph neural networks (Limbeck et al., 2026); dataset quality/curation (Couch et al., 2024) with implications for network distillation; and single- and multi-objective optimization (Huntsman, 2022, 2023a,b; Pereverdieva et al., 2025; Emmerich, 2026), and even chemistry (Bi et al., 2024, 2025). In short, many aspects of machine learning could benefit from techniques built on weighting vectors that were computed more scalably.

Our contributions are i) a systematic empirical analysis of scaling approaches for solving $Zw = 1$; ii) a containerized, open-source C++/MPI pipeline built on STRUMPACK (Rouet et al., 2016) at <https://github.com/Cynnovative/compass>, and iii) documented failures of other hitherto plausible scaling approaches that may be useful to other researchers.

The paper is organized as follows. §2 gives background on magnitude, weightings, and diversity; §3 discusses approaches to scaling; §4 details our methods and experiments; and §5 discusses scaling. Appendices §A and §B respectively give examples illustrating failures of sparsification and Nyström approximations; and §C details various semianalytical results that may be of interest but fail to produce obvious algorithmic advantage.

2. Weightings, magnitude, and diversity

We now get into details. A square matrix $Z \geq 0$ with $\text{diag}(Z) > 0$ is called a *similarity matrix*. We consider similarity matrices of the form $Z = \exp[-td]$ where $(f[M])_{jk} := f(M_{jk})$ indicates componentwise function application, $t \in (0, \infty)$, and d is a square matrix with

1. For $t \in (0, t_+)$, the diversity-maximizing distribution can be **NP**-hard to compute, since it requires consideration of exponentially many sparsity patterns. Moreover, while in Euclidean space there is a polynomial-time algorithm for $t = 0$ (Huntsman, 2026a), it still requires computing $d^{-1}1$ for a series of distance submatrices, which is at least as hard as computing $Z^{-1}1$ for $t > 0$.

entries in $[0, \infty]$ that also satisfy the triangle inequality. In this paper, we only consider Euclidean metric matrices of the form $d_{jk} = \|x_{(j)} - x_{(k)}\|$ for $\{x_{(j)}\}_{j=1}^n \subset \mathbb{R}^m$.

A *weighting* w is a solution to

$$Zw = 1, \quad (1)$$

where 1 indicates a vector of all ones. If Z has a weighting w , its *magnitude* is $\text{Mag}(Z) := 1^T w$. If d is Euclidean, then Z is positive definite and it has a unique weighting. Weightings turn out to be excellent scale-dependent boundary or outlier detectors in Euclidean space (Willerton, 2009; Bunch et al., 2020) due to a link with Bessel capacities (Meckes, 2015).

Example 1 Let $\{x_j\}_{j=1}^3 \subset \mathbb{R}^2$ with $d_{jk} := d(x_j, x_k)$ given by $d_{12} = d_{13} = 1 = d_{21} = d_{31}$ and $d_{23} = \delta = d_{32}$ with $\delta \ll 1$. A brief calculation shows that

$$w_1 = \frac{e^{(\delta+2)t} - 2e^{(\delta+1)t} + e^{2t}}{e^{(\delta+2)t} - 2e^{\delta t} + e^{2t}}; \quad w_2 = w_3 = \frac{e^{(\delta+2)t} - e^{(\delta+1)t}}{e^{(\delta+2)t} - 2e^{\delta t} + e^{2t}}.$$

As Figure 2 shows, for $t \ll 1$, $w \approx (1/4, 1/4, 1/2)^T$; for $t \gg 1$, $w \approx (1, 1, 1)^T$, and for $t \approx 10$, $w \approx (1/2, 1/2, 1)^T$. Thus the “effective number of points” goes from $\approx 1/4 + 1/4 + 1/2 = 1$, to $\approx 1/2 + 1/2 + 1 = 2$, to $\approx 1 + 1 + 1 = 3$.

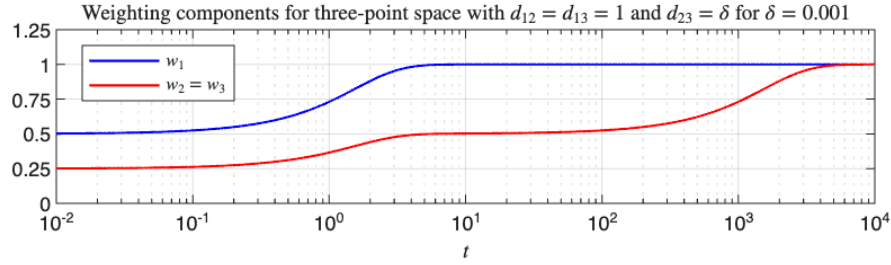


Figure 2: The magnitude $w_1 + w_2 + w_3$ is a scale-dependent “effective number of points.”

Magnitude and weightings play a key role *maximizing* a general and axiomatically defined measure of diversity (Leinster and Meckes, 2016; Leinster, 2021). For $1 < q < \infty$, the *diversity of order q* for a probability distribution p and similarity matrix Z is defined as

$$\log D_q^Z(p) := \frac{1}{1-q} \log \sum_{j:p_j>0} p_j (Zp)_j^{q-1} \quad (2)$$

and extended via limits for $q = 1, \infty$. $D_q^Z(p)$ is a “correct” measure of diversity in very much the same way that Shannon entropy is a “correct” measure of information. In fact, $\log D_q^Z(p)$ is a geometrical generalization of the Rényi entropy of order q . The usual Rényi entropy corresponds to $Z = I$, and Shannon entropy corresponds to $Z = I$ and $q = 1$.

Theorem 1 If Z is symmetric, positive definite, and has a unique positive weighting w , then for all q , w is proportional to $\arg \max_p D_q^Z(p)$ (Leinster and Meckes, 2016).

Theorem 1 reduces diversity maximization to linear algebra while rendering q irrelevant. Accordingly, the practically canonical nonzero scale for computing weightings is the *cutoff* t_+ , i.e., the least value of t such that $w \geq 0$: see Figure 1. The cutoff can be computed efficiently via root-finding for $\min_j w_j$ using the elementary bound $t_+ \leq \log(n-1)/\min_j \min_{k \neq j} d_{jk}$. In practice Ridders’ method (Ridders, 1979; Press et al., 2007) performs better than bisection due to its prioritization of low function evaluation count rather than low iteration count.²

The cutoff t_+ is the least scale that permits the application of Theorem 1. However, in practice the cutoff scale is often quite large in the sense that the resulting weighting has many components with values close to unity, limiting its utility. It is often more advantageous to work in the limit $t = 0$, though the corresponding weightings are often sparsely supported. This limit is examined in Huntsman (2026a) and Huntsman (2026b). Here, we focus on finding t_+ , though structured/kernel solvers can also be used for the $t = 0$ case.

3. Approaches to scaling

There are several promising approaches to solve $Zw = 1$ more efficiently at scale: e.g., enforcing sparsity, or using linear solvers for structured kernel matrices. §B and §C respectively detail failures of Nyström approximations and semianalytical techniques.

3.1. Sparsification

The obvious approach to sparsity is to threshold the similarity matrix Z , implicitly setting small entries to zero, say by efficient approximate K -nearest neighbor search using hierarchical navigable small world graphs (Malkov and Yashunin, 2018). However, the theory in its present state crucially requires Z to be positive definite to have useful guarantees about any information-theoretical interpretation of w . While this is guaranteed for a Euclidean distance matrix by classical results, thresholding Z usually destroys positive definiteness (Guillot and Rajaratnam, 2015; Guillot et al., 2016). While taking t large enough ensures positive definiteness even after thresholding, approximation errors still result, and in practice we want t as close to t_+ as possible, which in turn requires solving $Zw = 1$ accurately.

Meanwhile, “distance concentration” often effectively makes different pairwise distances indiscernible in high dimension. The probability distribution of distance $d_{12} = r$ between two IID points $\sim \mathcal{N}(0, I_m)$ rapidly approximates $\mathcal{N}(\sqrt{2(m-1)}, 1)$, as Figure 3 shows.³ While the essentially constant width of the distribution produces a narrowing of normalized distances, Gaussian tail behavior also raises the prospect that a quasilinear number of point pairs might control Z . However, numerical and analytical examinations of Gaussian-distributed data show that sparsification is not a tenable route to fast and accurate solutions of $Zw = 1$ at scale. Even as a first step, any effective sparsification via thresholding would

-
2. It is useful to halt the root-finder when $\min_j w_j \in (0, \varepsilon)$ for ε comparable to machine epsilon. This has the additional advantage of producing nondegenerate maximum-diversity probability distributions.
 3. To derive the distribution in Figure 3, note that the difference of two IID $\mathcal{N}(0, I_m)$ points is distributed as $\mathcal{N}(0, 2I_m)$: the norm of this difference is chi-distributed (in fact, $\chi_m(r/\sqrt{2})$). A mechanical derivation is in (Thiery and Hickman, 2015), and a more sophisticated partial generalization to $\mathcal{N}(0, \Sigma)$ with $\Sigma_{jj} \equiv 1$ is in (Martirosyan and Ohanyan, 2024). Meanwhile, a formula for the distribution of distances between two independent uniformly sampled points from a convex body is in (Aharonyan and Khalatyan, 2020).

require the identification of a small number of so-called “hubs” with many reverse nearest neighbors (Angiulli, 2018), which is also challenging. But in any event our numerical results from sparsification were so unsatisfactory as to not warrant significant discussion: see §A.

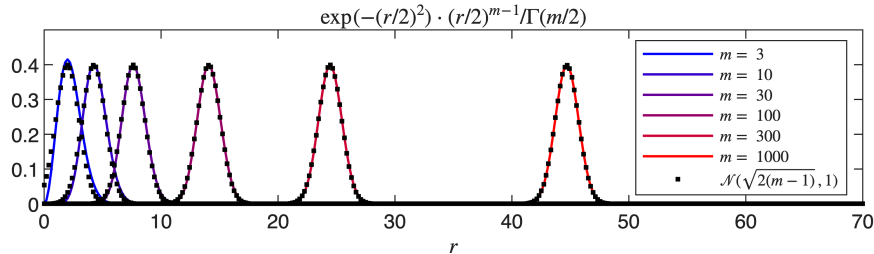


Figure 3: The distribution of distances between two IID points $\sim \mathcal{N}(0, I_m)$.

3.2. Solvers for structured kernel matrices

If d is a Euclidean distance matrix, then the corresponding Z is automatically positive definite, so we get a reproducing kernel Hilbert space to which Mercer’s theorem applies. In practice this amounts to the existence of a Cholesky decomposition. But a more promising method to solve $Zw = 1$ at scale is to use efficient hierarchical low-rank solvers that exploit kernel structure in a more detailed way and offer subquadratic performance. However, like the fast multipole method that spawned them, hierarchical solvers traditionally scale very badly with ambient problem dimension, or at best with data dimension (March et al., 2015).

Our experiments show that these techniques can offer real advantage: while dense kernel solvers scale well with dimension, hierarchical solvers permit much larger scale applications, albeit with dimension dependence.

4. Methods, experiments, and results

Our numerical experiments were performed on a 16-core MacBook Pro M4 Max with 128 GB RAM and ARM CPU (the integrated GPU was not exercised since it did not offer CUDA). We implemented two main computational pipelines: a high-performance C++/MPI back-end for large kernel matrix experiments using STRUMPACK (Rouet et al., 2016), and a Python-based pipeline for dense and sparse approaches. All experiments are conducted in Docker containers to ensure reproducibility of environment configurations. Docker images are configured to run on either x86 or ARM CPUs. We made all 16 CPUs available for Docker and capped its memory at 66.5 GB. Thread and device environments (OMP_NUM_THREADS, OPENBLAS_NUM_THREADS etc.) are set in the appropriate containers.

4.1. Software environment

C++/STRUMPACK pipeline. For large-scale experiments involving hierarchical low-rank approximation and kernel solves, we developed a C++ driver compiled against the STRUMPACK library (v7.2.0), with MPI parallelism enabled and GPU support. The current build uses OpenBLAS (v0.3.21) and OpenMPI for BLAS/LAPACK and parallel sup-

port, with optional CUDA (12.3+) for GPU acceleration. SLATE (v2025.05, for distributed BLAS) and METIS (v5.1) were also linked for fast matrix reordering and factorization.

Python/SciPy pipeline. For dense solving, we used `scipy.linalg.solve` in Python 3.12, with multithreaded BLAS operations supported by OpenBLAS. All package versions are managed using `micromamba` and are specified via YAML file.

Docker configurations. Our C++ pipeline is based on `nvidia/cuda:12.3.1-devel-ubuntu22.04` to enable leveraging CUDA GPU in the future. Our Python pipeline is based on `mambaorg/micromamba:2.3.1` to ensure we can link SciPy and NumPy to the most performant linear algebra libraries (e.g., OpenBLAS) and leverage multi-threaded processing.

4.2. Experimental workflow

For both the C++ and Python pipelines, the key algorithm is root-finding for t_+ , as described above. As a sanity check that adds marginal time, we test positive (semi)definiteness by estimating the smallest eigenvalue (C++: symmetric Lanczos via STRUMPACK; Python: `scipy.sparse.linalg.eigsh`). To speed up the Python pipeline, we first perform a fast Cholesky check with `scipy.linalg.cholesky` and only perform an eigenvalue decomposition if this check fails. We solve (1) using (`strumpack::HSSMatrix::solve`) in C++ and `scipy.linalg.solve` in Python.

All code, build recipes, environment files, and Docker images (for both C++ and Python stacks) are in GitLab and will be publicly released to ensure full reproducibility of all results.

4.3. Experimental results

To demonstrate the pipeline and gauge runtimes, we sampled n points from $\mathcal{N}(0, I_m)$ and from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{jk} := \delta_{jk}2^{-(j-1)}$. We take $n \in \{1000, 3000, 10000, 30000\}$ and $m \in \{3, 10, 30, 100\}$, and we perform three runs per configuration.⁴ For each set of points, the computational task is to compute t_+ and the weighting at that scale, as discussed above.

In preliminary experiments not detailed here, sparsification resulted in unacceptable performance, with Kendall tau coefficients relative to exact results overwhelmingly below 0.85. Also, on relatively small problems, iterative methods (either through SciPy or STRUMPACK) solved as fast or faster than the dense pipeline, with similar levels of fidelity, but scaling behavior suggested that iterative methods are only marginally beneficial.

Figures 4-7 show that HSS methods through STRUMPACK scale better than ordinary dense methods from SciPy (even with BLAS optimization and multithreading): both runtime and memory use are noticeably better for the former for $n \gtrsim 10^4$, with significantly lower CPU utilization above low dimension. Because STRUMPACK is built for distributed computation, it is reasonable to visually extrapolate our time and memory results, provided that runtime is appropriately amortized over nodes. Notably, although HSS performance degrades with dimension (as expected), this degradation is more graceful than a catastrophic impact that might have been expected.

4. In general, the computational complexity of solving is at least linear in m because of the underlying kernel evaluations. If the intrinsic dimension of the data is fixed while the ambient dimension m grows, hierarchical solvers will generally continue to perform well (March et al., 2015).

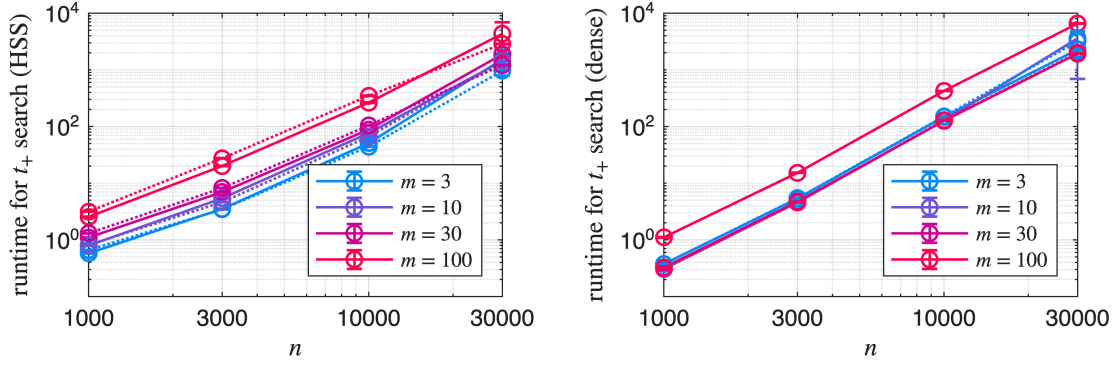


Figure 4: Left: total runtime (in seconds) for t_+ search using HSS methods. Results for $\Sigma = I_m$ are shown as solid lines; results for $\Sigma_{jk} := \delta_{jk} 2^{-(j-1)}$ are shown as dashed lines. Error bars are obtained from three experiments for each pair (m, n) . Right: as in the left panel, but for a dense solver. Not shown: the number of search iterations for $m = 100$ was much higher for the dense solver.

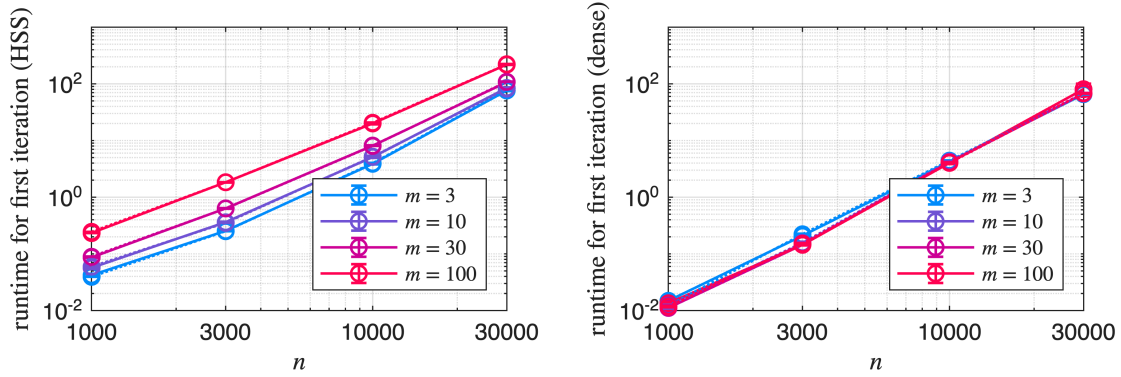


Figure 5: As in Figure 4, but for runtime (in seconds) for the first iteration of t_+ search.

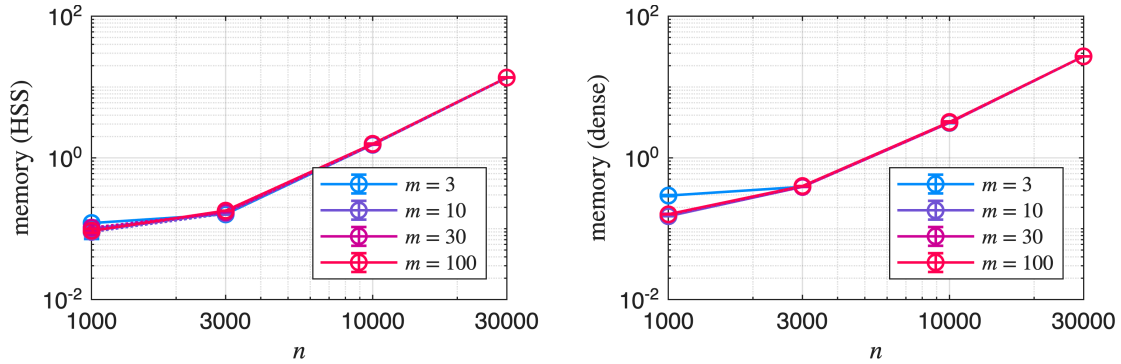


Figure 6: As in Figure 4, but for memory usage (in GB).

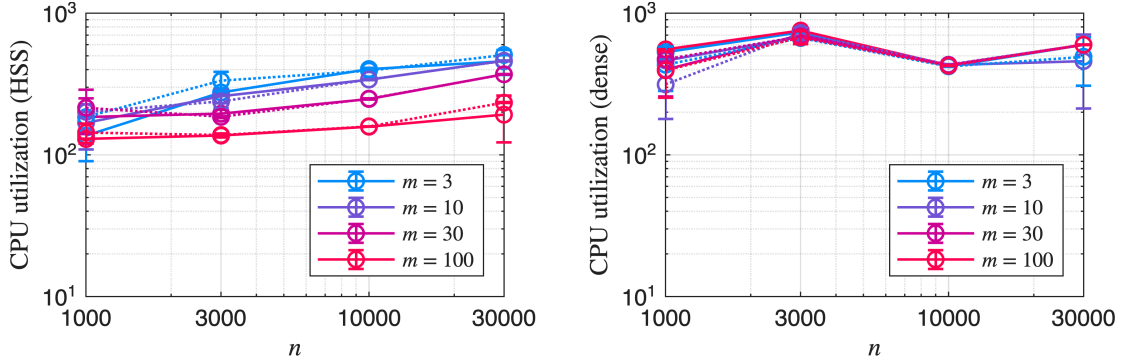


Figure 7: As in Figure 4, but for CPU utilization percentage.

5. Scaling

5.1. Current scaling

Our experiments deliberately focused on laptop-scale performance: however, more scaling is clearly possible using distributed MPI on a multi-node cluster and/or CUDA (Wang et al., 2019). We have implemented this and tested it, albeit on a single machine without CUDA, where overhead hinders performance. However, this is the correct approach for large problems on a multi-node cluster where MPI ranks have access to additional cores and memory bandwidth. Our Dockerfile builds with CUDA support, so access to NVIDIA hardware can provide GPU acceleration as well, primarily due to an internal *ULV* factorization involving dense block operations on the HSS tree.

For $n \approx 30000$, the time for a single iteration of the rootfinding algorithm (including but not just solving a single instance of $Zw = 1$) with STRUMPACK is typically on the order of one minute, which is comparable to MATLAB.⁵ Matrix compression takes about a quarter of that time; Lanczos slightly under ten percent; and *ULV* factorization takes about two thirds of the time. Around $n \approx 70000$, our solver segfaults inside OpenBLAS's `dgemm_itscopy` during compression, as a debugger backtrace confirms. This is an issue with BLAS, not STRUMPACK, and the practical ceiling for our current build is $n = 65000$. The underlying cause is likely large HSS ranks and/or compression tolerances.

We attempted various optimizations with little or no effect. For example, reducing the number of Lanczos iterations from 100 to 30 provided real but marginal performance gains. Loose HSS tolerances led to unacceptable noise, and we avoided them. Relaxing compression tolerances does not meaningfully reduce HSS compression ranks, because Gaussians are hard to compress. Replacing scattered heap allocations (`std::vector<std::vector<double>>`) with a single array improved cache locality and let the compiler vectorize distance computation with SIMD instructions, which accelerated compression by several percent, but compression itself is only about a quarter of the overall runtime.

5. There is no way to reuse factorizations of $Z(t)$ for different values of t , which informs the problem of finding the cutoff t_+ .

5.2. Future scaling

Any frontal assault on large-scale instances of (1) requires STRUMPACK or a functional analogue (e.g., GOFMM (Yu et al., 2017)) built for supercomputers. While scales of $n \approx 10^5$ are achievable on powerful workstations, substantially larger scales require clusters at $n \approx 10^6$, or leading supercomputing resources at $n \approx 10^8$.

Unfortunately, faster exact solutions or good approximations are not clearly in evidence, but many practical applications may not require either. For example, a black-box global optimizer that uses weighting vectors under the hood (Huntsman, 2022; Hoffman and Huntsman, 2022) might not actually require precise solutions of (1) in order to efficiently explore a function landscape. Good local approximations and/or bad global approximations would be very scalable and might work well enough even for hundreds of thousands or millions of points in hundreds or thousands of dimensions. We leave this, and particularly the investigation of alternative low-rank approximations, for future work.

Acknowledgments

Thanks to Evan Gorman for many useful conversations; and to Yang Liu for providing advice regarding STRUMPACK; and to referees for suggestions that helped the presentation.

We used Claude Opus to help find references and evaluate implementations.

This research was partially developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

- Amirhesam Abedsoltan, Parthe Pandit, Luis Rademacher, and Mikhail Belkin. On the Nyström approximation for preconditioning in kernel machines. In *International Conference on Artificial Intelligence and Statistics*, pages 3718–3726. PMLR, 2024.
- Michael F Adamer, Edward De Brouwer, Leslie O’Bray, and Bastian Rieck. The magnitude vector of images. *Journal of Applied and Computational Topology*, 8(3):447–473, 2024.
- NG Aharonyan and V Khalatyan. Distribution of the distance between two random points in a body from $\mathbb{R}^n$. *Journal of Contemporary Mathematical Analysis (Armenian Academy of Sciences)*, 55(6):329–334, 2020.
- Gernot Akemann, Jinho Baik, and Philippe Di Francesco. *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, 2011.
- Arash A Amini and Zahra S Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, 2021.
- Rayna Andreeva, Katharina Limbeck, Bastian Rieck, and Rik Sarkar. Metric space magnitude and generalisation in neural networks. In *Topological, Algebraic and Geometric Learning Workshops 2023*, 2023.

- Rayna Andreeva, James Ward, Primož Skraba, Jie Gao, and Rik Sarkar. Approximating metric magnitude of point sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15374–15381, 2025.
- Fabrizio Angiulli. On the behavior of intrinsically high-dimensional spaces: distances, direct and reverse nearest neighbors, and hubness. *Journal of Machine Learning Research*, 18(170):1–60, 2018.
- Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.
- Wanying Bi, Xin Fu, Jingyan Li, and Jie Wu. Persistent magnitude for the quantitative analysis of the structure and stability of carboranes. *Journal of Computational Biophysics and Chemistry*, 23(06):741–751, 2024.
- Wanying Bi, Hongsong Feng, Jie Wu, Jingyan Li, and Guo-Wei Wei. Topological magnitude for protein flexibility analysis. *Royal Society Open Science*, 12(12), 2025.
- Charles Bordenave. On Euclidean random matrices in high dimension. *Electronic Communications in Probability*, 18:1–8, 2013.
- Eric Bunch, Daniel Dickinson, Jeffery Kline, and Glenn Fung. Practical applications of metric space magnitude and weighting vectors. *arXiv preprint arXiv:2006.14063*, 2020.
- Eric Bunch, Jeffery Kline, Daniel Dickinson, Suhaas Bhat, and Glenn Fung. Weighting vectors for machine learning: numerical harmonic analysis applied to boundary detection. *arXiv preprint arXiv:2106.00827*, 2021.
- Difeng Cai, James Nagy, and Yuanzhe Xi. Fast deterministic approximation of symmetric indefinite kernel matrices with high dimensional datasets. *SIAM Journal on Matrix Analysis and Applications*, 43(2):1003–1028, 2022.
- Gustavo Chávez, Yang Liu, Pieter Ghysels, Xiaoye Sherry Li, and Elizaveta Rebrova. Scalable and memory-efficient kernel ridge regression. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 956–965. IEEE, 2020.
- Josiah Couch, Rima Arnaout, and Ramy Arnaout. Beyond size and class balance: alpha as a new dataset quality metric for deep learning. *arXiv preprint arXiv:2407.15724*, 2024.
- Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.
- Michael Emmerich. The magnitude of dominated sets: a Pareto compliant indicator grounded in metric geometry. *arXiv preprint arXiv:2604.18147*, 2026.
- Matthew Fickus and Dustin G Mixon. Tables of the existence of equiangular tight frames. *arXiv preprint arXiv:1504.00253*, 2015.
- Zachary Frangella, Joel A Tropp, and Madeleine Udell. Randomized Nyström preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2):718–752, 2023.

- Dominique Guillot and Bala Rajaratnam. Functions preserving positive definiteness for sparse matrices. *Transactions of the American Mathematical Society*, 367(1):627–649, 2015.
- Dominique Guillot, Apoorva Khare, and Bala Rajaratnam. Preserving positivity for matrices with sparsity constraints. *Transactions of the American Mathematical Society*, 368(12):8929–8953, 2016.
- Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Alaa Saade, Shantanu Thakoor, Bilal Piot, Bernardo Avila Pires, Michal Valko, Thomas Mesnard, Tor Lattimore, and Rémi Munos. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.
- Wolfgang Hackbusch. *Hierarchical Matrices: Algorithms and Analysis*. Springer, 2015.
- Zachary Hoffman and Steve Huntsman. Benchmarking an algorithm for expensive high-dimensional objectives on the BBOB and BBOB-largescale testbeds. In *Genetic and Evolutionary Computation Conference Companion*, 2022.
- Steve Huntsman. Parallel black-box optimization of expensive high-dimensional multimodal functions via magnitude. *arXiv preprint arXiv:2201.11677*, 2022.
- Steve Huntsman. Diversity enhancement via magnitude. In *International Conference on Evolutionary Multi-Criterion Optimization*, 2023a.
- Steve Huntsman. Quality-diversity in dissimilarity spaces. In *Genetic and Evolutionary Computation Conference*, 2023b.
- Steve Huntsman. Peeling metric spaces of strict negative type. In *Proceedings of the 1st Conference on Topology, Algebra, and Geometry in Data Science*, volume 321 of *Proceedings of Machine Learning Research*, 2026a.
- Steve Huntsman. Peel neighborhoods. *arXiv preprint arXiv:2603.26645*, 2026b.
- Nouredine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, pages 1–50, 2010.
- Tom Leinster. The magnitude of metric spaces. *Documenta Mathematica*, 18:857–905, 2013.
- Tom Leinster. *Entropy and Diversity*. Cambridge, 2021.
- Tom Leinster and Mark W Meckes. Maximizing diversity in biology and beyond. *Entropy*, 18(3):88, 2016.
- Remigijus Leipus, Jonas Šiaulys, Mantas Dirma, and Romualdas Zovė. On the distribution-tail behaviour of the product of normal random variables. *Journal of Inequalities and Applications*, 2023(1):32, 2023.
- Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. Metric space magnitude for evaluating the diversity of latent representations. *Neural Information Processing Systems*, 2024.

- Katharina Limbeck, Lydia Mezrag, Guy Wolf, and Bastian Rieck. Geometry-aware edge pooling for graph neural networks. In *Advances in Neural Information Processing Systems*. 2026.
- Yu. A Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2018.
- William B March, Bo Xiao, Sameer Tharakan, Chenhan D Yu, and George Biros. A kernel-independent FMM in general dimensions. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015.
- DM Martirosyan and VK Ohanyan. On the Euclidean distance between two Gaussian points and the normal covariogram of $\mathbb{R}^d$. *Journal of Contemporary Mathematical Analysis (Armenian Academy of Sciences)*, 59(1):38–46, 2024.
- Mark W Meckes. Magnitude, diversity, capacities, and dimensions of metric spaces. *Potential Analysis*, 42:549–572, 2015.
- Hans Z Munthe-Kaas, Olivier Verdier, and Gilles Vilmart. A short proof of Isserlis’ theorem. *arXiv preprint arXiv:2503.01588*, 2025.
- Ksenia Pereverdieva, André Deutz, Tessa Ezendam, Thomas Bäck, Hèrm Hofmeyer, and Michael Emmerich. Comparative analysis of indicators for multi-objective diversity optimization. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 58–71. Springer, 2025.
- William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes, 3rd edition*. Cambridge, 2007.
- C Ridders. A new algorithm for computing a single root of a real continuous function. *IEEE Transactions on Circuits and Systems*, 26(11):979–980, 1979.
- François-Henry Rouet, Xiaoye S. Li, Pieter Ghysels, and Artem Napov. A distributed-memory package for dense hierarchically semi-separable matrix computations using randomization. *ACM Transactions on Mathematical Software (TOMS)*, 42(4):1–35, 2016.
- Melvin Dale Springer and WE Thompson. The distribution of products of independent random variables. *SIAM Journal on Applied Mathematics*, 14(3):511–526, 1966.
- Željka Stojanac, Daniel Suess, and Martin Kliesch. On products of Gaussian random variables. *arXiv preprint arXiv:1711.10516*, 2017.
- Benjamin Thirey and Randal Hickman. Distribution of Euclidean distances between randomly distributed Gaussian points in n -space. *arXiv preprint arXiv:1508.02238*, 2015.
- Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems*, 2019.

Simon Willerton. Heuristic and computer calculations for the magnitude of metric spaces. *arXiv preprint arXiv:0910.5500*, 2009.

Chenhan D Yu, James Levitt, Severin Reiz, and George Biros. Geometry-oblivious FMM for compressing dense SPD matrices. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2017.

Appendix A. Examples of sparsification for Gaussian-distributed points

Figure 8 illustrates how sparsification using K approximate nearest neighbors obtained as in (Malkov and Yashunin, 2018) (and still symmetrizing) severely degrades solutions to $Zw = 1$ and to the calculation of cutoffs for Gaussian-distributed points. The bad approximations shown are broadly typical, though much worse results occurred frequently in other realizations, while materially better results did not occur.

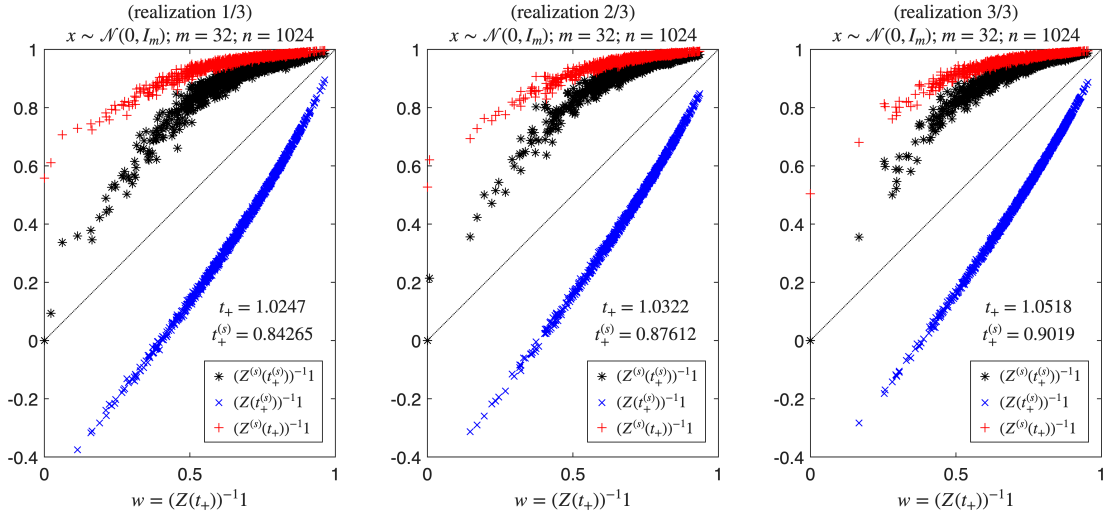


Figure 8: Effects of sparsification for three realizations of a sample of size $n = 1024$ from $\mathcal{N}(0, I_m)$ with dimension $m = 32$. For each sample, we compute the true cutoff t_+ and corresponding similarity matrix Z , as well as a symmetrized sparse approximation $Z^{(s)}$ using $\lceil \log_2(n) \rceil = 10$ approximate nearest neighbors and approximate cutoff $t_+^{(s)}$ computed using the same binary search technique as for t_+ . We show weighting approximations using the sparse cutoff (blue $\times$ markers), the sparse similarity matrix (red $+$ markers), and both simultaneously (black $*$ markers).

It may be surprising that sparsification has numerical effects as profound as Figure 8 indicates. However, it is unsurprising in light of §3.1 that sparsification destroys many practical theoretical guarantees as well.

Appendix B. Nyström approximation

The Nyström approximation (Cai et al., 2022) for a kernel matrix is of the form

$$\begin{pmatrix} K_{(LL)} & K_{(LM)} \\ K_{(LM)}^T & K_{(MM)} \end{pmatrix} \approx \begin{pmatrix} K_{(LL)} & K_{(LM)} \\ K_{(LM)}^T & K_{(LM)}^+ K_{(LL)} K_{(LM)} \end{pmatrix} \quad (3)$$

where a partition of points into a small “landmark” set L and a large “main” set M is indicated. It is reasonable to hope that this might help scalably approximate $Z^{-1}\mathbf{1}$ and $d^{-1}\mathbf{1}$.⁶ However, naive applications are essentially unusable for our (or similar) purposes that demand precision in order to compute t_+ and the associated weighting.

The most immediate reason is that the kernels $\exp(-t\|x - x'\|)$ and $\|x - x'\|$ generically produce matrices without any good low-rank approximation (cf. Figures 15 and 16). This leads to a numerical catastrophe. For the exponential kernel, this is easily mitigated by subtracting the diagonal of the right hand side of (3) and adding an identity matrix. However, this mitigation is not just obviously imperfect, but it also leaves some room for improvement.

Specifically, the resulting approximate weighting vectors tend to behave particularly badly at landmark points. Using uniformly random landmarks and performing affine transformations on the landmark and main components to match their first two moments while holding the main variance fixed yields an approximate weighting vector that is highly correlated with the exact result for a wide range of scales. However, the approximation is only quantitatively good at one scale, and it is not even clear how to find this scale *a priori*.

Using a greedy stochastic algorithm to find more “outlying” landmarks along the lines of Algorithm 2 in (Huntsman, 2023b) requires a different approach for transforming landmark and main components. In this event, it is more appropriate to transform the main components as above, but to transform the landmarks differently, by matching them with the upper tail of the main components, then shifting the landmark components by the average difference between this matching transformation and the one above.

This more careful application of Nyström approximations holds more promise than a naive application, but it still has no obvious utility. Figures 9 and 10 show how these approximations are strongly correlated with exact results, but with uncertain linear fits that strongly depend on t . Unfortunately, even if we can use this approximation to reliably identify the index of the minimal weighting component, it is not clear how to use that information to compute the cutoff scale more efficiently. For example, the intercepts of linear fits when $t = 0.4$ are far from zero, while the actual cutoff scale is just under 0.7 (cf. Figures 13 and 14).

Appendix C. Semianalytical investigations

C.1. Weightings at scale zero for Gaussian-distributed points

In practice, weightings at scale zero are often more useful than weightings at finite nonzero scale (Huntsman, 2026a). While the latter requires solving $Zw = \mathbf{1}$ once for $Z = \exp[-td]$

6. Although Nyström approximations are useful for preconditioning (Frangella et al., 2023; Abedsoltan et al., 2024), we do not consider this in detail here since STRUMPACK has robust preconditioning facilities.

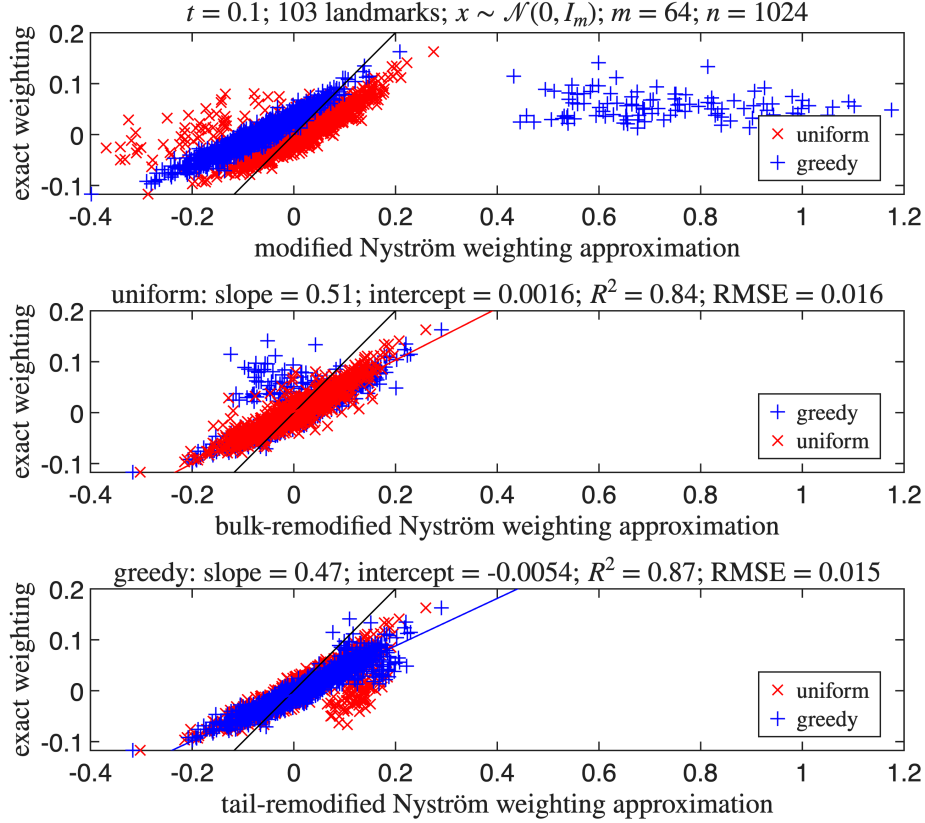


Figure 9: A comparison of modified Nyström approximations to a weighting using **uniform** and **stochastic magnitude-greedy** landmarks with $t = 0.1$. The upper panel compares the approximations obtained with a unit diagonal to the exact result, shown as the black diagonal line. In each case, the outliers are the landmarks. The middle panel shows the results of affine transformations on landmark and main components by basic moment-matching: this works better for uniform landmarks: a corresponding linear fit is also shown. The lower panel shows the results of a more delicate transformation suitable for greedy landmarks: a corresponding linear fit is also shown.

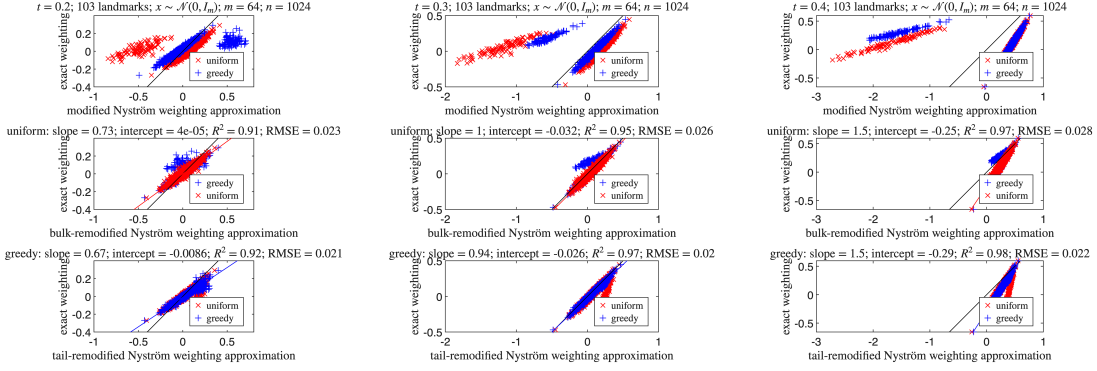


Figure 10: As in Figure 9, but for $t = 0.2$ (left), $t = 0.3$ (center), and $t = 0.4$ (right).

and d a Euclidean (or more generally here, strict negative type) distance matrix, the former requires solving equations of the form $dw = 1$, since

$$\lim_{t \downarrow 0} (\exp[-td])^{-1} \mathbf{1} = \frac{d^{-1} \mathbf{1}}{\mathbf{1}^T d^{-1} \mathbf{1}}. \quad (4)$$

To get diversity-maximizing distributions, it is necessary to solve equations of the form $dw = 1$ repeatedly, albeit at progressively smaller scales.

Meanwhile, a “canonically hard” case for solving the weighting equation at zero or finite scale is presented by a sample of n IID points $x^{(j)}$ from a standard m -dimensional Gaussian, i.e., $x^{(j)} \sim \mathcal{N}(0, I_m)$ for $j \in [n]$.

For the scale zero case and recalling the approximation $\mathcal{N}(\sqrt{2(m-1)}, 1)$ of the distribution of distances between IID points $\sim \mathcal{N}(0, I_m)$, we have that the linear equation $dw = 1$ is approximated by the equation

$$(A + \sqrt{2(m-1)}(\mathbf{1}\mathbf{1}^* - I))w = 1, \quad (5)$$

where $A_{jk} \sim \mathcal{N}(0, 1)$ for $j \neq k$ and $A_{jj} \equiv 0$.⁷ An IID instance of (5) is a so-called Wigner ensemble (Akemann et al., 2011), but to our knowledge the structure of solutions to $dw = 1$ has not been studied. Our case of course has entries that are obviously not independent even if approximately identically distributed.

As we shall see, the structure of solutions to an instance of (5) with IID upper triangular entries of A differs dramatically from (1) for Gaussian-distributed points. But let us entertain the alternative hope briefly. As justification, it is reasonable to try to characterize and compare solutions to the Wigner ensemble and actual scale zero weighting equations (for which entries are not independent) with the same entrywise statistics to try to develop a heuristic for sparsification. Because the standard Gaussian mass is concentrated in a thin spherical shell, the distance matrix behaves particularly badly from the point of view of sparsification, and any heuristic that worked effectively in this situation would be likely to be applicable very generally.⁸

7. The case for finite scale is obviously better behaved, but harder to work with analytically and also less useful: $Z_{jk} \sim \text{LogNormal}(-t\sqrt{2(m-1)}(1 - \delta_{jk}), t^2)$.

8. A stringent test of such a heuristic would be an overcomplete frame that is approximately equiangular and tight, say by perturbing overcomplete equiangular tight frames such as those enumerated in (Fickus

Applying the Sherman-Morrison formula to (5) generically yields

$$\left(A + \sqrt{2(m-1)}(11^* - I)\right)^{-1} 1 = \frac{\left(A - \sqrt{2(m-1)}I\right)^{-1} 1}{1 + 1^* \left(A - \sqrt{2(m-1)}I\right)^{-1} 1} \quad (6)$$

and

$$\begin{aligned} \left(A - \sqrt{2(m-1)}I\right)^{-1} 1 &= -\frac{1}{\sqrt{2(m-1)}} \left(I - \frac{1}{\sqrt{2(m-1)}}A\right)^{-1} 1 \\ &\approx -\frac{1}{\sqrt{2(m-1)}} \left(I + \frac{1}{\sqrt{2(m-1)}}A\right) 1. \end{aligned} \quad (7)$$

In the IID case, $A1 \sim \mathcal{N}(0, (n-1)I)$, so the maximum-diversity distribution is approximately uniform plus a Gaussian perturbation.⁹ The approximation (7) could conceivably not be as bad as it looks: while the entries of the left hand side are more often negative than not, the denominator of the factor multiplying it is distributed as $\mathcal{N}(1 - n/\sqrt{2(m-1)}, n(n-1)/4m^2)$, and $1 - n/\sqrt{2(m-1)}$ will typically be negative in practice.

In any event, numerics are required to get a practical understanding and to see if this model has any use at all. And it turns out that numerics quickly dispel the hope entertained above. Figure 11 illustrates (kernel density estimates of) the entries of d and A along the lines of (5), while Figure 12 illustrates how the IID approximation (5) to $d^{-1}1$ and its further approximation (6) are very bad.

C.2. Cutoff scales for Gaussian-distributed points

Intriguingly, a numerical experiment suggests that for $x^{(j)} \sim \mathcal{N}(0, I_m)$ with $j \in [n]$, we have the excellent approximation

$$\mathbb{E}t_+ \approx \gamma_+ m^{-2/3} n^{1/4} \quad (8)$$

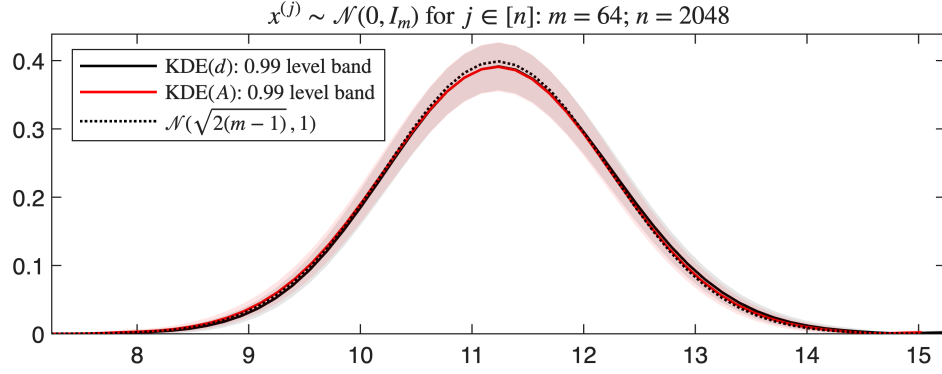
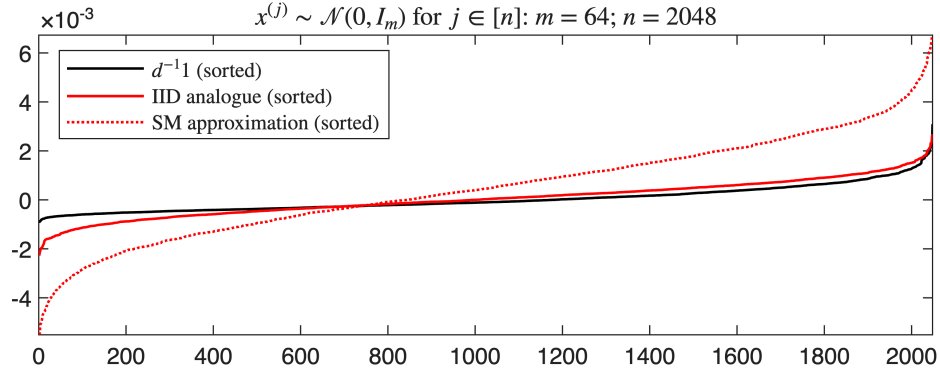
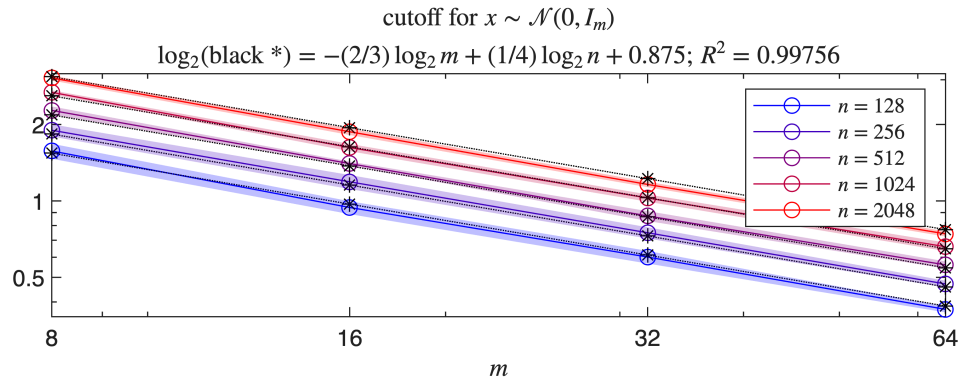
where γ_+ is slightly less than 2. Figures 14 and 13 show the results of computing cutoffs for n points distributed as $\mathcal{N}(0, I_m)$ for varying dimension m and sample size n .

and [Mixon, 2015](#)). This would arguably be the worst possible case to confront in practice, though the prospects for characterization are too weak to warrant starting here.

9. The next higher order correction to (7) is governed under the IID *Ansatz* by

$$\begin{aligned} (A^2 1)_j &= \sum_{k\ell} A_{jk} A_{k\ell} \\ &= \sum_{k, \ell \neq j} A_{jk} A_{k\ell} + \sum_k A_{jk}^2 \\ &\sim (\pi^{-1} K_0(|\cdot|))^{*(n-1)^2} * \chi_{n-1}^2 \end{aligned}$$

and in principle (perhaps even practice) this can be evaluated further. Higher-order corrections are yet more complicated. Even the expectations of $A^\ell 1$ under the IID *Ansatz* involve Isserlis' theorem ([Munthe-Kaas et al., 2025](#)) as a building block. Meanwhile, the actual products of IID Gaussian random variables are distributed as Meijer G -functions ([Springer and Thompson, 1966](#); [Stojanac et al., 2017](#)) whose tail behavior has been analyzed in ([Leipus et al., 2023](#)).


 Figure 11: Density estimates for d and A along the lines of (5).

 Figure 12: Sorted entries of $d^{-1}1$, the IID analogue (5), and a subsequent approximation to (6) using (7).

 Figure 13: Cutoffs of $N = 10$ realizations of size n from $\mathcal{N}(0, I_m)$, with standard deviations indicated, and with (8) in black for comparison (taking $\gamma_+ \approx 2^{7/8} \approx 1.834$). For smaller m and n the linear behavior breaks down and the standard deviations are much larger.

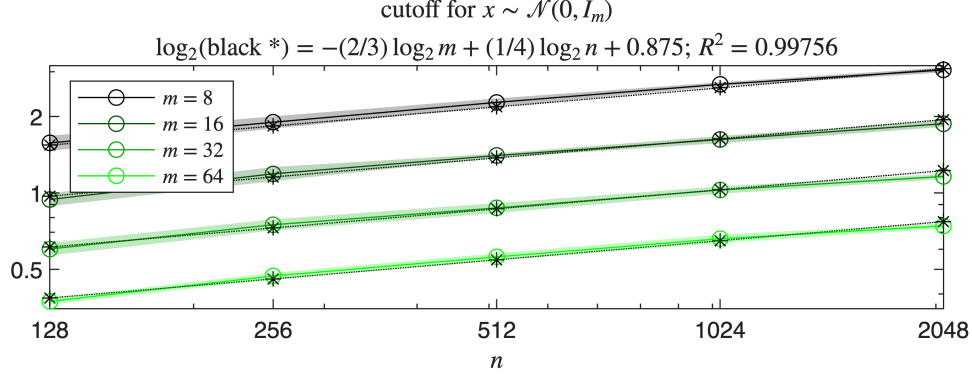


Figure 14: As in Figure 13, but as a function of sample size n vs. of dimension m .

It is natural to try to exploit (8) to compute actual cutoffs more quickly. One idea along these lines is to use (8) to produce an initial guess for a cutoff that might be better than the easy upper bound. The trivial scaling relationship $t_+(\lambda d) = \lambda^{-1} t_+(d)$ indicates how to normalize d towards this end. However, this idea does not seem to save (or cost) much time, and it is not clear how to exploit (8) in practice.

Still, (8) is sufficiently intriguing to warrant further investigation.

C.3. Random kernel matrices

Let $x^{(j)} \sim \mathcal{N}(0, m^{-1} I_m)$ be IID for $j \in [n]$, and as usual let $d_{jk} = \|x^{(j)} - x^{(k)}\|$ be the corresponding distance matrix. This is a special case of a *Euclidean random matrix* (Karoui, 2010; Bordenave, 2013), or more generally, a *random kernel matrix* (Do and Vu, 2013) of the form

$$M_{jk} := F(x^{(j)}, x^{(k)}, m). \quad (9)$$

Random kernel matrices concentrate around their mean (Amini and Razaei, 2021), and as such they tend to have bad numerical behavior: similarity matrices are no exception here.

On the other hand, Theorem 2 of (Do and Vu, 2013) gives the nice result that for

$$F(x^{(j)}, x^{(k)}, m) = f(\|x^{(j)} - x^{(k)}\|^2)$$

with f differentiable at 2, M has the same limiting spectral distribution as

$$B := [f(0) - f(2) + 2f'(2)] I_n - 2f'(2) X^T X, \quad (10)$$

where $X_{j\ell} := x_\ell^{(j)}$. Since $\text{spec}(\alpha I + \beta A) = \alpha + \beta \cdot \text{spec}(A)$, the spectral distribution of (10) follows trivially from the spectral distribution for the Wishart ensemble $X^T X$, which tends to a Marchenko-Pastur distribution (Bai and Silverstein, 2010; Akemann et al., 2011).

In our case, we have $f_t(\xi) := \exp(-t\sqrt{\xi})$, and example spectra of $Z(t_+)$ and B are shown in Figure 15. For $t = 0$, we have instead $f(\xi) = \sqrt{\xi}$, but the same theorem still applies, as shown in Figure 16.

Unfortunately, it is not clear if or how we can usefully leverage knowledge of the limiting spectral distribution in general. For example, there is probably little value in constructing

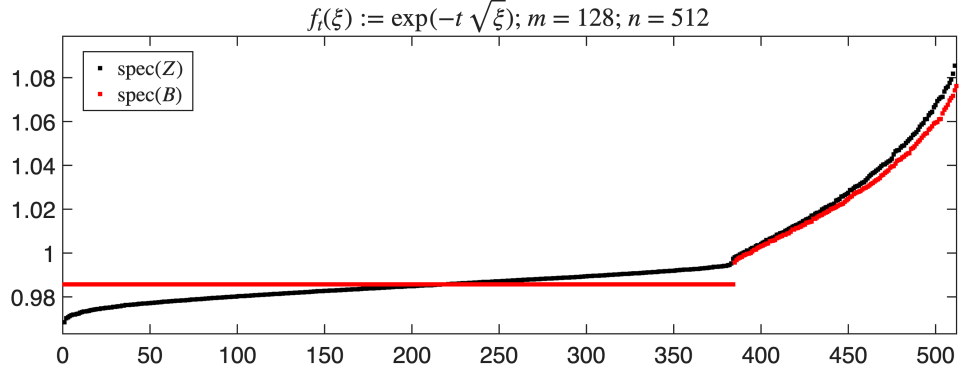


Figure 15: The spectrum of $Z(t_+)$ (black) and the limiting spectral distribution from (10) (red). The largest eigenvalue ≈ 3.15 of $Z(t_+)$ is not shown.

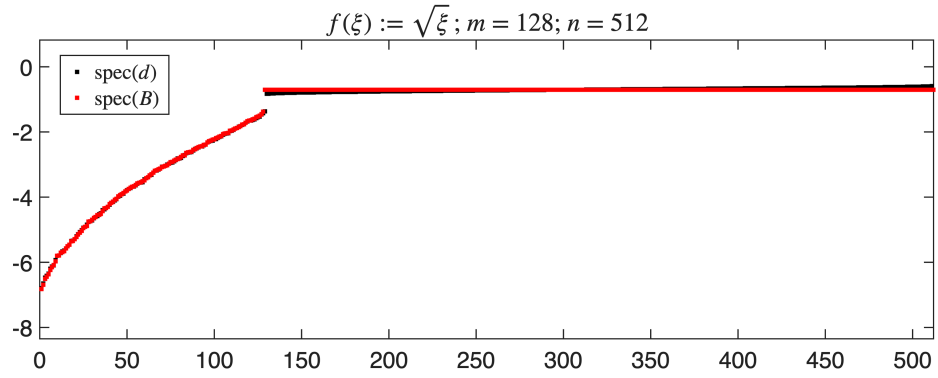


Figure 16: The spectrum of d (black) and the limiting spectral distribution from (10) (red). The largest eigenvalue ≈ 720 of d is not shown.

a preconditioner to accelerate computations on normally distributed data. While the relevant theorem also applies to normalized uniform samples from certain nice Riemannian submanifolds with the canonical metric inherited from $\mathbb{R}^m$ (e.g., spheres) (Karoui, 2010), the confluence of uniformity and metric requirements render this ostensibly broader applicability largely irrelevant in practice. Furthermore, convergence to the limiting distribution is poor unless $m \gg 1$, in which case the curse of dimensionality suggests that any sample suitable for the theorem would be hard to distinguish than a sample from the standard normal distribution.

C.4. Summary

The semianalytical approximations for scale zero weightings that we have considered all fail. While cutoff scales for isotropic Gaussian-distributed points are easy to approximate using (8) and the relation $t_+(\lambda \cdot d) = \lambda^{-1} \cdot t_+(d)$, the utility of this approximation appears to be insignificant in practice. Similarly, it is not clear if how to obtain an estimate of the spectrum of Z or of d outside of situations where its utility is basically zero (e.g., a preconditioner for normally distributed data). In short, semianalytical techniques appear to have little or no practical value for solving the weighting equation, even approximately.