Executive summary

In the 21st century, we have witnessed major increases in data availability and computational resources, yet algorithms for data analysis remain painfully slow. Chemists struggle to find clusters in data sets with $N > 10^5$ molecular configurations because spectral clustering would require factorizing an $N \times N$ matrix, which is too large to even store on a computer. Meanwhile, geophysicists struggle to simulate Category 5 hurricanes reaching Boston, because generating high-resolution simulations would require thousands of computers operating for years. Motivated by such seemingly impossible computations, my role as an applied mathematician is to build new randomized algorithms that offer speed-up factors of at least $10 \times -100 \times$ and support these algorithms by proving mathematical guarantees. By providing much-needed and well-justified randomized algorithms, I enable scientists to perform calculations that would otherwise be intractable. Meanwhile, this work enables me to discover the probabilistic and linear algebraic structures that make randomized algorithms effective.

As my first research program, I am using randomized low-rank matrix approximation to speed up kernel machine learning. Traditionally, kernel machine learning requires $O(N^3)$ floating point operations, which restricts the applicability to data sets with $N \leq 10^5$ points. However, I am developing a new randomized strategy that reduces the cost from $O(N^3)$ to $O(k^2 N)$ operations where $k$ is the matrix approximation rank [Che+23; Díá+23]. Before it was common to run kernel spectral clustering codes for over a day on high-performance workstations [Roh+11]. In contrast, my approach accurately clusters $N = 2.5 \times 10^5$ data points in just 10 seconds on a laptop [Che+23]. To mathematically support this approach, I have proved new error bounds [TW23; Che+23] that quantify the close relationship between approximation accuracy and kernel matrix eigenvalue decay.

Second, I am developing a new rare event sampling algorithm inspired by the survival of the fittest. The algorithm duplicates the “fittest” simulations (deemed likely to lead to the rare event) and randomly eliminates the “least fit” simulations (deemed unlikely to lead to the rare event). The algorithm leads to a $7 \times$ speed-up when simulating intense hurricanes near Boston [Web+19] and a $30 \times$ speed-up when simulating Mercury’s close encounters with Venus [Abb+21]. I have proved the algorithm produces unbiased estimates [Web+19] and am working to establish optimality. This research has been featured in Forbes.com, the California Business Journal, and SIAM News.

Below, I describe my research programs in more detail and state my overall research vision.

Program 1: Accelerate kernel machine learning

Kernel methods are a popular set of algorithms for prediction and clustering. Kernel methods perform interpretable AI, since they use an explicit kernel function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to quantify the similarity between points $x_1, \ldots, x_N \in \mathbb{R}^d$. The strong kernel hypothesis posits that kernel methods can compete with machine learning algorithms based on neural networks [Rad+22; Lee+20; Aro+20], but it is currently impossible to certify or refute this hypothesis because kernel methods are too costly to be applied to the largest data sets. My vision is to resolve the strong kernel hypothesis by building the computational tools needed to deploy kernel methods at scale.
The computational bottleneck in a kernel method is a linear system of equations or an eigenvalue problem involving the kernel matrix $K \in \mathbb{R}^{N \times N}$ with entries $k_{ij} = \phi(x_i, x_j)$. Even storing the kernel matrix is difficult with $N > 10^5$ data points, and traditional matrix inversion or eigendecomposition requires an exorbitant $O(N^3)$ floating point operations. As a cheaper alternative, my approach replaces the kernel matrix $K$ with a randomized low-rank approximation $\hat{K}$, thus reducing the cost to $O(Nk^2)$ operations where $k$ is the selected approximation rank. Figure 1 shows that a rank $k = 150$ leads to nearly perfect kernel spectral clustering when applied to a biochemistry data set with $N = 2.5 \times 10^5$ points. For this example, the difference between $N^3$ and $Nk^2$ is a factor of 3 million, and the method runs in ten seconds on a laptop [Che+23].

![Figure 1](image)

Figure 1: Randomized kernel matrix approximation leads to near-perfect clustering of chemistry data (left), revealing 2-dimensional clusters in a 30-dimensional data space (right). See [Che+23].

To explain the success of randomized kernel matrix approximation, I have proved new error bounds. Let $[K]_r$ denote the theoretically optimal rank-$r$ approximation of a positive semidefinite matrix $K \in \mathbb{R}^{N \times N}$, from an $r$-truncated eigendecomposition. I have proved that randomized block Krylov iteration [TW23, Thm. 7.1] generates a low-rank approximation $\hat{K}$ that satisfies

$$
\mathbb{E}\| K - \hat{K}\| \leq (1 + \varepsilon)\| K - [K]_r\|, \quad \text{for } r \geq 1, \quad \varepsilon > 0,
$$

using an approximation rank $k = \tilde{O}(r\varepsilon^{-1/2})$, where the $\tilde{O}$ notation suppresses logarithmic factors. Similarly, I have proved that randomly pivoted Cholesky [Che+23, Thm. 3.1] satisfies

$$
\mathbb{E}\text{tr}(K - \hat{K}) \leq (1 + \varepsilon)\text{tr}(K - [K]_r), \quad \text{for } r \geq 1, \quad \varepsilon > 0,
$$

using a larger approximation rank $k = \tilde{O}(r\varepsilon^{-1})$. Bounds for randomized block Krylov iteration were known before my work [MM15], but I improved these bounds to be numerically accurate and contain explicit constants. My bounds for randomly pivoted Cholesky are completely new.

**Next steps.** The efficiency of kernel machine learning can potentially be improved even further, to just $O(k^3 + kN)$ operations and $O(k^2 + N)$ storage. Such a low cost is possible because randomized low-rank approximation can be implemented in just $O(k^3)$ operations with rejection sampling [EM23] and memory requirements can be minimized with stochastic gradient descent [MB17; MB19; ABP23]. Over the next two years, I will combine these ingredients to form an accelerated strategy with error bounds building on my earlier techniques [Che+23; Día+23].

---

**Program 2: Accelerate rare event probability calculation**

Extreme weather events, such as hurricanes, floods, and heat waves, are traditionally studied using repeated simulations from a high-resolution model. Yet, a rare event with probability $p$ occurs just

---

Robert J. Webber

✉️ rwebber@caltech.edu • 🔍 rwebber.people.caltech.edu
once per $p^{-1}$ simulations. Generating enough samples of the rare event might require a thousand or ten thousand simulations, and each high-resolution simulation can require weeks of computing.

To prevent these exploding computing costs, I introduced a rare event sampling algorithm that makes replicas of the “fittest” simulations that are leading to a rare event, while randomly eliminating many of the “least fit” simulations [Web+19]. I proved the algorithm yields unbiased estimates and established a variance formula to enable a posteriori error estimation [Abb+21]. Figure 2 shows an example in which the algorithm (right) produces several samples of Category 5 hurricanes near Boston (probability $p = 0.002$), while direct simulation (left) produces none.

Next steps. I am advising a team of planetary scientists that is using the survival-of-the-fittest algorithm to evaluate Mercury’s probability of experiencing a close encounter with Venus [Abb+21; Abb+23a; Abb+23b]. Meanwhile, I am developing a theory to explain the optimal variance that any Monte Carlo rare event sampling scheme can achieve [WAS22; Ari+23], and I plan to extend this theory to establish an optimality guarantee for the survival-of-the-fittest algorithm.

Research vision

In the 21st century, we are transitioning from deterministic algorithms, which often run slowly, to modern randomized algorithms that require less memory and operations. During this transition period, new mathematical insights are needed to ensure the accuracy of the randomized algorithms. My research has already shown how randomized algorithms exploit low-rank structure [TW23; Che+23] and sparse structure [WW23] to provide fast and accurate solvers. My vision is to expand this library of exploitable structures further, making randomized methods so easy to use and effective that they are taught in every graduate math and science curriculum.

As an example of a future research direction, I will use iterative random sparsification to accelerate large-scale fixed-point calculations arising in physics and chemistry. In this approach, I replace a convergent but computationally expensive fixed-point iteration $x_{t+1} = F(x_t)$ with a cheaper iteration $x_{t+1} = F(y_t)$, where $y_t = \text{sparse}(x_t)$ is a sparse randomized approximation. Related techniques have been used to solve eigenvalue problems as large as $10^{108} \times 10^{108}$ in quantum chemistry [She+12], as well as infinite-dimensional linear systems in PDE analysis [CDD01], but without guaranteed accuracy. By studying these techniques mathematically, I will identify opportunities to exploit randomness in a principled way, leading to faster and more reliable calculations.

I am also excited to investigate active learning algorithms, which identify a small representative subset of a large data set $x^{(1)}, \ldots, x^{(N)} \in \mathbb{R}^d$. The most powerful active learning algorithms

Robert J. Webber

✉️ rwebber@caltech.edu • • rwebber.people.caltech.edu
[Mus+22] currently require expensive pre-processing steps, involving a complete pass over the data. Yet I perceive the possibility of a high-performing online active learning algorithm, where the user is presented with random data points and accepts or rejects each data point with probability depending on the previous selections. An early version of this approach was suggested in [EM23], but I see the potential to take many fewer samples while achieving higher accuracy.

In conclusion, randomness is needed now more than ever to advance the computational frontier. As a faculty member, I will develop new randomized algorithms which address the needs of scientists struggling with large-scale computations. Meanwhile, I will benefit from these interactions with scientists, using their computational challenges as a springboard for new developments in linear algebra and probability theory.

References


Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. “Active Linear Regression for $\ell_p$ Norms and Beyond”. In: IEEE 63rd Annual Symposium on Foundations of Computer Science. 2022. DOI: 10.1109/FOCS54457.2022.00076.

Adityanarayanan Radhakrishnan, George Stefanakis, Mikhail Belkin, and Caroline Uhler. “Simple, fast, and flexible framework for matrix completion with infinite width neural networks”. In: Proceedings of the National Academy of Sciences 119.16 (2022), e2115064119. DOI: 10.1073/pnas.2115064119.


