Sample complexity bounds for localized sketching

Rakshith S Srinivasa
Georgia Institute of Technology

Mark A. Davenport
Georgia Institute of Technology

Justin Romberg
Georgia Institute of Technology

Abstract

We consider sketched approximate matrix multiplication and ridge regression in the novel setting of localized sketching, where at any given point, only part of the data matrix is available. This corresponds to a block diagonal structure on the sketching matrix. We show that, under mild conditions, block diagonal sketching matrices require only $O(\mathbf{sr}/\epsilon^2)$ and $O(\mathbf{sd}_\lambda/\epsilon)$ total sample complexity for matrix multiplication and ridge regression, respectively. This matches the state-of-the-art bounds that are obtained using global sketching matrices. The localized nature of sketching considered allows for different parts of the data matrix to be sketched independently and hence is more amenable to computation in distributed and streaming settings and results in a smaller memory and computational footprint.

1 Introduction

Efficient linear algebraic computations are of fundamental importance in machine learning and signal processing applications. This has led to a rise in randomized linear algebraic methods that aim to solve large problems only approximately, but with much less time complexity compared to standard methods (see [Woodruff, 2014, Wang et al., 2017, Yang et al., 2016, Chowdhury et al., 2018] and references therein). In this work, we consider two specific examples: sketched matrix multiplication [Cohen et al., 2015] and ridge regression [Avron et al., 2016] but with additional constraints on the sketching matrices that arise in the context of distributed data acquisition. Formally, if $\mathbf{W} \in \mathbb{R}^{\widetilde{N}\times m}$ and $\mathbf{Y} \in \mathbb{R}^{\widetilde{N}\times p}$, computing the product $\mathbf{W}^T\mathbf{Y}$ takes $O(mp\widetilde{N})$ time, which can be prohibitive for large $\widetilde{N}$. The sketched version then aims to find matrices $\mathbf{S} \in \mathbb{R}^{\widetilde{M}\times\widetilde{N}}$ such that

$$\left\|(\mathbf{SW})^T(\mathbf{SY}) - \mathbf{W}^T\mathbf{Y}\right\| \leq \epsilon \left\|\mathbf{W}\right\| \left\|\mathbf{Y}\right\|. \quad (1)$$

Computing the sketched matrix product $(\mathbf{SW})^T(\mathbf{SY})$ then takes only $O(mp\widetilde{M})$ time (not accounting the time to compute $\mathbf{SW}$ and $\mathbf{SY}$ themselves). State-of-the-art bounds show that $\widetilde{M} = O(\max(\mathbf{sr}(\mathbf{W}), \mathbf{sr}(\mathbf{Y}))/\epsilon^2)$ suffices, where $\mathbf{sr}(\cdot)$ is the stable rank of a matrix (defined in Section 2 and is a stable alternative for the rank). Similarly, given $\mathbf{A} \in \mathbb{R}^{\widetilde{N}\times d}$ with $\widetilde{N} \gg d$ and $\mathbf{b} \in \mathbb{R}^{\widetilde{N}}$, the ridge regression problem is

$$\mathbf{x}_* = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min}\ f(\mathbf{x}) := \left\|\mathbf{Ax} - \mathbf{b}\right\|^2 + \lambda \left\|\mathbf{x}\right\|^2 \quad (2)$$

and can be solved in $O(\widetilde{N}d^2)$ time. The sketched problem instead seeks to find matrices $\mathbf{S} \in \mathbb{R}^{\widetilde{M}\times\widetilde{N}}$ such that solving

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min}\ f_{\mathbf{S}}(\mathbf{x}) := \left\|\mathbf{SAx} - \mathbf{Sb}\right\|^2 + \lambda \left\|\mathbf{x}\right\|^2 \quad (3)$$

yields

$$f(\hat{\mathbf{x}}) \leq (1+\epsilon) f(\mathbf{x}_*). \quad (4)$$

The state-of-the-art bounds show that for small ϵ, $\widetilde{M} = O(\mathbf{sd}_\lambda/\epsilon)$ suffices, where $\mathbf{sd}_\lambda$ is the statistical dimension and is again a more stable alternative to the rank of $\mathbf{A}$, as defined in Section 2.

With this background in place, let us consider a scenario where the data matrix $\mathbf{A}$ is naturally divided into J blocks that are not all available at a single location. Let each block then be of size $N \times d$, where $\widetilde{N} = JN$. Such partitioning of data into different blocks occurs naturally in many applications. For example, dynamic systems produce data that evolve over time. To store the entire data before sketching it would require large amounts of memory [9]. It would be of use to sketch the system as it evolves, leading to a natural partition. In yet another application, consider the square kilometer array [11]. This array consists of antennas distributed across the continents of Australia and Africa. To handle the massive data rates (157 TB/s), it is desirable to

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

sketch the data locally at each antenna and then transmit to the central processing location. In distributed systems that use edge-cloud architecture, edge nodes collect data that needs to be communicated to the cloud for inference. The communication requirements can be made smaller if the data at each edge node is compressed to an "optimal" dimension.

A feature of existing sketching methods (including those that use fast Johnson-Lindenstrauss matrices such as Subsampled Randomized Hadamard Transform (SRHT) [Ailon and Chazelle, 2006] and sparse sketching matrices [Clarkson and Woodruff, 2017]) is that they need access to all or an arbitrary subset of the rows of $\mathbf{A}$ (See Figure 1). Clearly, this is unsuitable for an application with distributed data. This leads us to ask the following questions: Is there a way to adapt sketching techniques to such applications? What is the best way to model dimensionality reduction for such applications? Two naïve ways are readily available: i) Since each block is of size $N \times d$, its rank is upper bounded by d. One could obtain a subspace embedding for each block and communicate these sketched blocks to the central node. The resulting dimension of the aggregated data is then $O(Jd/\epsilon^2)$, since each block needs to be sketched to $O(d/\epsilon^2)$, ii) Sketch each data block separately, and **add** the resulting sketches at the central node instead of aggregating them. In fact, this results in a sketch of the entire data matrix $\mathbf{A}$. Using existing bounds, one can conclude that the final sketch needs to be $O(d/\epsilon^2)$, which again requires each data block also to be sketched to $O(d/\epsilon^2)$.

A major drawback of both of the above approaches is that they do not take advantage of the inherent low dimensionality of the entire matrix $\mathbf{A}$, resulting in a sketch size of $O(d/\epsilon^2)$ for each data block. Our observation is that it should be possible to lose information locally, while still retaining all the information about $\mathbf{A}$ globally. This is exactly what we address in this paper: we show theoretically that it is possible for the each of the blocks to be sketched to $O(d/J\epsilon^2)$. This implies that the sketch obtained from a single block may not be big enough to provide a subspace embedding for that block. Yet, an embedding of the entire matrix $\mathbf{A}$ can be obtained, once the sketches from the individual blocks are aggregated. Hence, our work aims to initiate a study of how to extend sketching methods to distributed data acquisition scenarios.

Our proposal is to impose a block diagonal structure on the sketching matrix $\mathbf{S}$. We denote such a sketching matrix as $\mathbf{S}_D$. We then partition the data matrices $\mathbf{W}$, $\mathbf{Y}$ and $\mathbf{A}$ analogously. This results in sketches of the form

$$\mathbf{S}_D\mathbf{A} = \begin{bmatrix} \mathbf{S}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{S}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{S}_J \end{bmatrix} \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_J \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1\mathbf{A}_1 \\ \mathbf{S}_2\mathbf{A}_2 \\ \vdots \\ \mathbf{S}_J\mathbf{A}_J \end{bmatrix}. \tag{5}$$

We assume that $\mathbf{A}_j \in \mathbb{R}^{N\times d}$ where $\widetilde{N} = JN$ and $\mathbf{S}_j \in \mathbb{R}^{M_j \times N}$ such that $\sum_j M_j = \widetilde{M}$, although our results extend to the case where the $\mathbf{A}_j$'s are of different sizes. Further, in our paper we assume that the nonzero entries of the matrix $\mathbf{S}_D$ are drawn from the Gaussian distribution. Our goal is to study the sample complexities $\widetilde{M}_j$ required to achieve similar guarantees as those in [Cohen et al., 2015] and [Avron et al., 2016] for dense (non-block diagonal) sketching matrices.

Apart from the structural advantages described above, computing the product $\mathbf{S}_D\mathbf{A}$ can also be much cheaper when compared to an unstructured random projection. For generic $\mathbf{S}_j$, the sketch $\mathbf{S}_D\mathbf{A}$ can be computed in time $O(Nd\widetilde{M})$, as compared to the $O(\widetilde{N}d\widetilde{M})$ required for a dense, unstructured sketch. Second, the computation is trivial to parallelize into J blocks, each requiring $O(NdM_j)$ time. For large problems with low effective rank, when we can take $M_j = O(\log N)$, this gives us a sketch with structured randomness competitive with methods that use SRHT and sparse embedding matrices [Woodruff, 2014]. Furthermore, the blocks themselves could be designed to be fast transforms. Owing to these computational advantages, blocking could be a strategy by itself.

1.1 Related work

There is a vast and growing literature on sketching techniques. Here we briefly review some of the work most relevant to ours in the context of our setting. Note that while sketching can also be used as a preconditioning method [Yang et al., 2016], here we will only address "sketch and solve" methods where the original problem is (approximately) solved in a reduced dimension.

Sketching methods for solving ordinary least squares problems are well summarized in [Woodruff, 2014]. However, as noted in [Avron et al., 2016], solutions for sketched ridge regression problems are more relevant in practice since regularization is often necessary. Similar to [Avron et al., 2016], we address this problem but in the setting where the sketching matrix is block diagonal. We provide conditions on the matrix $\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$ under which such structured matrices can have the same sample complexity as [Avron et al., 2016].

Our work is closely related to that of

[Eftekhari et al., 2015] which studies the restricted isometry property (RIP) of block diagonal matrices. These results can be used to directly obtain subspace embedding guarantees for block diagonal matrices. However, this approach requires a sample complexity dependent on the rank of $\mathbf{A}$ and not its approximate rank. For large matrices with fast spectral decay, this dependency can lead to sub-optimal sample complexity. Another difference is that we consider block diagonal matrices that have different sized blocks, while [Eftekhari et al., 2015] assumes that all the blocks are of the same size. One of the main conclusions of our paper is that choosing the block sizes in a data dependent fashion leads to improved (optimal) sample complexity.

A statistical analysis of sketched ridge regression in a distributed setting is provided in [Wang et al., 2017]. This work considers the ridge regression problem in the multivariate setting (where $\mathbf{b}$ and $\mathbf{x}$ are matrices) and analyzes model averaging in the case of distributed computation of the sketched ridge regression solution. In this setting, various processors each solve the problem with a part of the data and the estimators are then communicated to a central agent. In contrast, we consider a scenario where the estimate is computed by the central agent with only sketched data sent from various nodes.

Another work that is similar in spirit to ours and addresses sketched regression in a distributed setting is [McWilliams et al., 2014]. The setting considered in this work lies somewhere between that of [Wang et al., 2017] and ours. It considers multiple processors solving the ridge regression problem with different parts of the data similar to [Wang et al., 2017], but also assumes that the data used by each processor is available to all other processors in a sketched form. In contrast, in our work, the sketched data from all the nodes is available to only a central computing agent.

A complimentary line of work focuses on the same problem but where $\widetilde{N} \ll d$. In [Chen et al., 2015], a sketching based algorithm is proposed that achieves a relative error guarantee for the solution vector. This result is further improved in [Chowdhury et al., 2018]. Sketching has also been applied in the context of kernel ridge regression, where the data points are mapped to higher dimensional feature space before solving the regression problem. Sketching is used to reduce the number of such high dimensional features in [Paul and Drineas, 2016] and [Avron et al., 2017]. Sampling and rescaling of features is considered in [Paul and Drineas, 2016]. Random feature maps are also used to construct pre-conditioners in [Avron et al., 2017] to solve kernel ridge regression, where it is shown that a number of random feature maps proportional to the effective rank of the kernel matrix suffices to obtain a high quality pre-conditioner. While our work targets a different setting (where $\widetilde{N} \gg d$) and requires a different set of analytical tools, it is noteworthy that our guarantees involve a similar dependence on the stable rank of the underlying data matrix.

2 Main results

Our main contribution is theoretical analysis of the block model described in (5). A naïve strategy to analyze block diagonal matrices is to treat each block $\mathbf{A}_j$ separately and use a number of random projections proportional to its effective rank. But this would not take advantage of the low dimensional structure of the full matrix $\mathbf{A}$, resulting in a highly suboptimal sample complexity. Instead, we show that under mild assumptions on $\mathbf{A}$, the total sample complexity of $\widetilde{M}$ of the matrix $\mathbf{S}_D$ can match the existing bounds mentioned above.

2.1 Stable rank, statistical dimension and incoherence

Before we can state our main results, we need to define a few quantities that characterize the *complexity* of matrix multiplication and ridge regression problems.

Stable rank of a matrix: The stable rank of a matrix $\mathbf{W}$ is defined as $\mathbf{sr}(\mathbf{W}) = \frac{\|\mathbf{W}\|_F^2}{\|\mathbf{W}\|^2}$. Note that $\mathbf{sr}(\mathbf{W}) \leq \text{rank}(\mathbf{W})$. For matrices with a flat spectrum, the stable rank equals the rank of the matrix. However, if the singular values decay, then the stable rank captures the effective low dimensionality of the matrix, even when it is technically full rank.

Statistical dimension of the ridge regression problem: The ridge regression problem defined in (2) can be reformulated as

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\| \begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|^2 \Leftrightarrow \min_{\mathbf{x}\in\mathbb{R}^d} \left\| \widetilde{\mathbf{A}}\mathbf{x} - \widetilde{\mathbf{b}} \right\|^2.$$

The scalar multiple of the identity on the bottom of $\widetilde{\mathbf{A}}$ means it will technically be rank d. But in some sense, a more nuanced notion of rank would count dimensions in the column space of $\widetilde{\mathbf{A}}$ that have singular values greater than $\sqrt{\lambda}$ differently that those with singular values less than $\sqrt{\lambda}$. One way to make to bring this distinction out is through the *statistical dimension*

$$\mathbf{sd}_\lambda = \sum_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda}.$$

In the sum above, if $\sigma_i^2 \gg \lambda$, then the contribution for that term is approximately one, while if $\sigma^2 \ll \lambda$, it is essentially zero. This allows us to interpret $\mathbf{sd}_\lambda$ as a kind of "effective rank". Note that $\mathbf{sd}_\lambda \leq \text{rank}(\mathbf{A})$

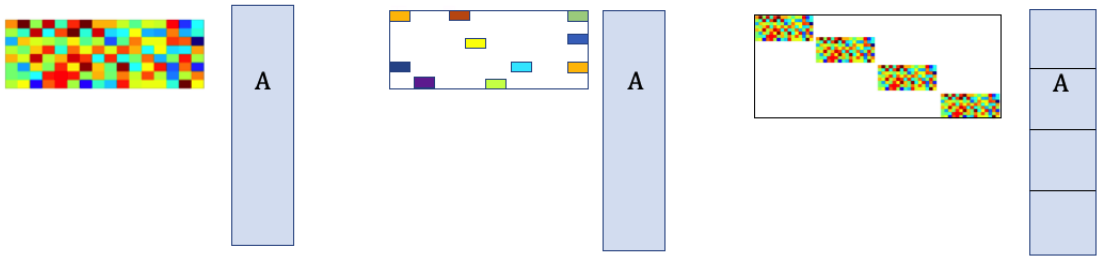

Figure 1: Existing sketching strategies such as dense sub-Gaussian, SRHT matrices (left) and sparse sketching matrices (center) assume access to all or a few arbitrarily placed rows of $\mathbf{A}$. However, our localized model (right) needs access to only well-separated parts of the data matrix.

and can be much lower than rank($\mathbf{A}$). While making λ very large can of course make $\mathbf{sd}_\lambda$ very small, this also introduces a larger bias in the estimates provided by (2) and (3), driving both of their solutions to zero. Choosing the λ that balances this bias-variance trade-off is equally important in sketched and non-sketched ridge regression.

Incoherence of the data matrices: In randomized sampling schemes, the sampling probability of each row depends on the corresponding *leverage score*, which is the ℓ_2 norm of the corresponding row of an orthobasis $\mathbf{U}$ for $\mathbf{A}$. Leverage scores highlight the relative importance of each row of $\mathbf{A}$.

Block diagonal matrices can be thought of as a generalization of sampling matrices. Instead of a single row, each block now accesses a submatrix of $\mathbf{A}$. Instead of using uniformly sized diagonal blocks $\mathbf{S}_j$, we show that a relative importance term associated with each block $\mathbf{A}_j$ similar to leverage scores dictates the number of random projections M_j required to attain optimal sample complexity. Let $\mathbf{U}$ be an orthobasis for the column space of the matrix $\mathbf{A}$. Let $\mathbf{U} = [\mathbf{U}_1^T\ \mathbf{U}_2^T\ \cdots\ \mathbf{U}_J^T]^T$, where $\mathbf{U}_j \in \mathbb{R}^{N\times d}$. We will show that the corresponding relative importance parameter, which we term as *coherence* of $\mathbf{U}_j$, is

$$\Gamma(\mathbf{U}_j) = \min\left(\|\mathbf{U}_j\|_\infty^2 N, \|\mathbf{U}_j\|_2^2\right).$$

Here, $\|\mathbf{U}_j\|_\infty$ denotes the element-wise infinity norm and $\|\mathbf{U}_j\|_2$ denotes the spectral norm. We can observe that

$$\frac{1}{J} \leq \max_j \Gamma(\mathbf{U}_j) \leq 1. \tag{6}$$

When the $\Gamma(\mathbf{U}_j)$'s are all close to $1/J$, the columns of $\mathbf{U}$ are incoherent, or not too aligned with respect to the standard basis vectors. On the contrary, when they are close to 1, then there are vectors in the column space of $\mathbf{U}$ which are close (in an inner product sense) to the standard basis vectors. We describe bases $\mathbf{U}$ that have small coherence parameters as being *incoherent*. We will show that as long as the coherence is not too high, the sample complexity of block diagonal matrices can match that of generic sketching matrices.

Number of random projections: Low values of the coherence parameter (highly incoherent bases) indicate relative uniformity in the importance of the blocks. For such subspaces, it would be reasonable to expect that roughly the same number of random projections can be drawn from each data block $\mathbf{A}_j$. On the other hand, when the coherence parameters $\Gamma(\mathbf{U}_j)$ have a high dynamic range, it can be expected that the number of random projections from each block should be proportional to the corresponding $\Gamma(\mathbf{U}_j)$. This is precisely our proposed strategy to design the number of random projections M_j. We propose that M_j can be chosen as

$$M_j = M_0\Gamma(\mathbf{U}_j) \tag{7}$$

for some constant M_0 that we will determine later. Our theoretical results state that block diagonal sketching matrices can achieve optimal sample complexity when M_j's are designed as in (7). This is also reminiscent of sampling algorithms, where the sampling probability of each row is proportional to the corresponding leverage score.

2.2 Sample complexity bounds for localized sketching

Localized sketching for matrix multiplication

Some of the earlier works that addressed this problem required $\mathbf{S}$ to be of size $\Omega\left(\frac{r(\mathbf{W})+r(\mathbf{Y})}{\epsilon^2}\right) \times \widetilde{N}$ where $r(\cdot)$ denotes the rank of the matrix. However, matrices with high ranks can still be approximately low dimensional, as indicated by their stable rank. In [Cohen et al., 2015] it is shown that the sample complexity of $\mathbf{S}$ in (1) (under certain distributions) depends only on the *stable ranks* of the matrices. They describe

distributions $\mathcal{D}$ that satisfy

$$\mathop{\mathbb{P}}_{\mathbf{S}\sim\mathcal{D}}\Big(\left\|(\mathbf{SW})^T(\mathbf{SY}) - \mathbf{W}^T\mathbf{Y}\right\| > \epsilon \left\|\mathbf{W}\right\| \left\|\mathbf{Y}\right\| \sqrt{(1+\mathbf{sr}(\mathbf{W})/k)}\sqrt{(1+\mathbf{sr}(\mathbf{Y})/k)}\Big) < \delta \quad (8)$$

for any desired k and a suitable $\widetilde{M}$. When $\mathbf{S}$ is a dense matrix with sub-Gaussian entries, this holds for $\widetilde{M} = \Omega(\frac{k+\log(1/\delta)}{\epsilon^2})$. Then, for $k = \max(\mathbf{sr}(\mathbf{W}),\ \mathbf{sr}(\mathbf{Y}))$, $\mathbf{S}$ satisfies (1). Hence, to achieve a relative error in the spectral norm, $\mathbf{S}$ only needs to have a number of rows that is proportional to the stable ranks of $\mathbf{W}$ and $\mathbf{Y}$.

Our first main result is such a guarantee for block diagonal sketching matrices. Unlike the distributions proposed in [Cohen et al., 2015], block diagonal distributions cannot be both oblivious to the data matrices and have optimal sample complexity. A naïve way to achieve (8) when $\mathbf{S}$ is block diagonal is to use triangle inequality:

$$\left\|(\mathbf{S}_D\mathbf{W})^T(\mathbf{S}_D\mathbf{Y}) - \mathbf{W}^T\mathbf{Y}\right\| \leq \sum_j \left\|(\mathbf{S}_j\mathbf{W}_j)^T(\mathbf{S}_j\mathbf{Y}_j) - \mathbf{W}_j^T\mathbf{Y}_j\right\|$$

where $\mathbf{W}_j$ and $\mathbf{Y}_j$ are corresponding blocks as in (5). However, this requires that $M_j = \Omega\left(\frac{\mathbf{sr}(\mathbf{W}_j)+\mathbf{sr}(\mathbf{Y}_j)}{\epsilon^2}\right)$ for each j. This can lead to suboptimal sample complexities, as $\mathbf{sr}(\mathbf{W}_j)$ and $\mathbf{sr}(\mathbf{Y}_j)$ can be as high as $\mathbf{sr}(\mathbf{W})$ and $\mathbf{sr}(\mathbf{Y})$ themselves. We show in our analysis that we can in fact achieve

$$\widetilde{M} = \sum_j M_j = \Omega\left(\frac{\mathbf{sr}(\mathbf{W}) + \mathbf{sr}(\mathbf{Y})}{\epsilon^2}\right)$$

for incoherent matrices. With M_j designed as in (7), we have the following result for computing approximate matrix products:

Theorem 1 *Fix matrices $\mathbf{W}$ and $\mathbf{Y}$ and let $\mathbf{S}_D$ be a block diagonal matrix as in (5) with the entries of $\mathbf{S}_j$ are drawn from the distribution $\mathcal{N}(0, 1/M_j)$. Let $\mathbf{U}$ be an orthobasis for the matrix $[\mathbf{W}\ \mathbf{Y}]$ and $\Gamma(\mathbf{U}_j)$ be the corresponding incoherence terms. Then the tail bound (8) holds with $\mathbf{S} = \mathbf{S}_D$ when M_j are taken as in (7) with*

$$M_0 = \Omega\left(\frac{k\log(2/\delta)}{\epsilon^2}\right). \quad (9)$$

We can examine the total sample complexity of $\mathbf{S}_D$. Consider a highly incoherent basis $\mathbf{U}$: each entry of such a basis is bounded away from 1. Examples of such bases include orthobases of matrices with entries drawn from the Gaussian distribution and any subset of the Fourier basis. Since each column of $\mathbf{U}$ has an ℓ_2-norm of 1, for such bases, $\left\|\mathbf{U}_j\right\|_\infty \approx 1/\sqrt{\widetilde{N}}$. Then we have $M_j \approx \frac{M_0}{J}$ and $\widetilde{M} = \Omega\left(\frac{\max(\mathbf{sr}(\mathbf{W}),\mathbf{sr}(\mathbf{Y}))\log(2/\delta)}{\epsilon^2}\right)$. We see that even though $\mathbf{S}_D$ has a block diagonal structure, it can still have an optimal sample complexity.

Block diagonal sketching of ridge regression

Let us now consider the sketched ridge regression problem shown in (3). Let $\mathbf{U}_1 \in \mathbb{R}^{\widetilde{M}\times d}$ comprise the first n rows of an orthobasis for the matrix $[\begin{smallmatrix}\mathbf{A}\\ \sqrt{\lambda}\mathbf{I}_d\end{smallmatrix}]$. Then, (4) holds with constant probability, if $\mathbf{S}$ satisfies the following two conditions:

$$\left\|\mathbf{U}_1^T\mathbf{S}^T\mathbf{S}\mathbf{U}_1 - \mathbf{U}_1^T\mathbf{U}_1\right\| \leq \frac{1}{4}, \quad (10)$$

$$\left\|\mathbf{U}_1^T\mathbf{S}^T\mathbf{S}\mathbf{r}^* - \mathbf{U}_1^T\mathbf{r}^*\right\| \leq \sqrt{\frac{\epsilon f(\mathbf{x}^*)}{2}}, \quad (11)$$

where $\mathbf{r}_* = \mathbf{b} - \mathbf{A}\mathbf{x}^*$ and we recall that $f(\mathbf{x}^*) = \left\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\right\|^2 + \lambda\left\|\mathbf{x}^*\right\|^2$. These conditions are well known in the randomized linear algebra community. (See [Avron et al., 2016] Lemma 9.) Both of the above conditions on $\mathbf{S}$ can be re-expressed as approximate matrix product guarantees by choosing the pair of matrices as $\mathbf{W} = \mathbf{Y} = \mathbf{U}_1$ for (10) and $\mathbf{W} = \mathbf{U}_1$ and $\mathbf{Y} = (\mathbf{b} - \mathbf{A}\mathbf{x}^*)$ for (11). We now state our main result for block diagonal sketching of ridge regression problems. Let $\mathbf{A}$ and $\mathbf{b}$ be as defined above and let $\mathbf{U}$ be an orthobasis for a basis for the range of $[\mathbf{A}\ \mathbf{b}]$ of size at most $\widetilde{N}\times(d+1)$ with $\Gamma(\mathbf{U}_j)$'s being the corresponding incoherence terms.

Theorem 2 *Let $\mathbf{U}$ be an orthobasis for the matrix $[\mathbf{A}\ \mathbf{b}]$ and $\Gamma(\mathbf{U}_j)$ be the corresponding incoherence terms. Let $\mathbf{S}_D$ be a block diagonal matrix as in (5) with the entries of $\mathbf{S}_j$ are drawn from the distribution $\mathcal{N}(0, 1/M_j)$. Let $\mathbf{x}_*$ be the solution to (2), and $\hat{\mathbf{x}}$ be the solution to (3). Then*

$$f(\hat{\mathbf{x}}) \leq (1+\epsilon)f(\mathbf{x}_*),$$

with constant probability when M_j obeys (7) with $M_0 = \Omega\left(\frac{\mathbf{sd}_\lambda}{\epsilon}\right)$.

As before, if $\mathbf{A}$ and $\mathbf{b}$ are such that the basis $\mathbf{U}$ is incoherent, then the total sample complexity $\widetilde{M} = \sum_j M_j = O(\frac{\mathbf{sd}_\lambda}{\epsilon})$. We are hence able to establish that though highly structured, block diagonal random matrices can in fact have optimal sample complexities.

Estimating the incoherence terms An important question is about how the coherence parameters $\Gamma(\mathbf{U}_j)$'s can be estimated. Note that the main challenge is in computing an orthobasis for the data matrix $\mathbf{A}$. We develop an algorithm to empirically estimate the $\Gamma(\mathbf{U}_j)$'s to within a constant factor of the true values using a sketching based algorithm. The algorithm uses $O(d)$ fast localized random projections of the blocks $\mathbf{A}_j$'s and computes an estimate of the QR factorization of $\mathbf{A}$ at a central processing unit. Using

the approximate R factor, the blocks $\mathbf{U}_j$'s are estimated locally. The algorithm is detailed in the supplementary material and has a worst case time complexity of $O(\widetilde{N}d\log N)$. Note that this is less than the sketch compute time $O(\widetilde{N}d\widetilde{M}/J)$ for N not too large. In Figure 3, we show the estimated incoherence parameters and the true parameters for a test matrix with $J = 100$, $\widetilde{N} = 10000$. We can see that the estimated values are within a constant factor of the true $\Gamma(\mathbf{U}_j)$'s. An important note here is that in many applications, an estimate of the $\Gamma(\mathbf{U}_j)$'s may be obtained using a priori domain knowledge. Yet another insight is that if distributional assumptions on the data can be made, as common in machine learning, then $\Gamma(\mathbf{U}_j)$'s can be very reliably estimated a priori [Eftekhari et al., 2015]. Any such prior information will lead to better sample complexities as compared to the naïve techniques described in the introduction.

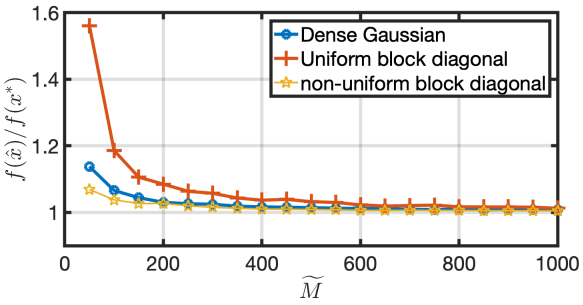

Figure 2: $f(\hat{x})/f(x^*)$ for three sketching matrices: a dense matrix with standard Gaussian entries, a block diagonal matrix with equal sized blocks (uniform diagonal matrix) and a block diagonal matrix with entries designed as in (7) (non-uniform diagonal matrix). A ratio close to 1 indicates that the sketching matrix is effective in solving (3). When M_j's are chosen appropriately, block diagonal matrices can be as effective as a general matrix.

3 Experiments

We demonstrate the effectiveness of block diagonal sketching matrices by performing experiments on both synthetic and real data. In our first experiment, we demonstrate the importance of choosing the size of the diagonal blocks according to our proposed method given in (7). We use the following parameters: $N = 2000$, $J = 10$, $d = 50$. We design the singular values such that for $\lambda = 0.15$, $\mathbf{sd}_\lambda = 8.5$, but $\text{rank}(\mathbf{A}) = 50$. For each trial, we generate $\mathbf{S}$ with entries drawn from $\mathcal{N}(0, 1/\sqrt{\widetilde{M}})$ and $\mathbf{S}_D$ with the entries of $\mathbf{S}_j$ drawn from $\mathcal{N}(0, 1/\sqrt{M_j})$. In Figure 2, we plot $f(\hat{x})/f(x^*)$ averaged over 10 trials for different values of $\widetilde{M}$. In particular, we show that when $M_j = M_0\Gamma(\mathbf{U}_j)$, $\mathbf{S}_D$ has the same rate of decay for $f(\hat{x})/f(x^*)$ as $\mathbf{S}$, and has a worse rate otherwise.

In our next set of experiments, we study performance in terms of prediction accuracy on the YearPrediction-MSD dataset. It contains 89 audio features of a set of songs and the task is to predict their release year. The dataset has 463,715 training samples and 51,630 test samples. In this case, we use diagonal blocks of the same size. Across 10 independent realizations of $\mathbf{S}$ and $\mathbf{S}_D$, we compute the empirical probability of $f(\hat{x})/f(x^*) \leq (1+\epsilon)$ for various values of ϵ and $\widetilde{M}$. We show phase transition plots in Figure 4 which demonstrate that block diagonal matrices are as effective as dense matrices in terms of accuracy, for the same sample complexity.

We also seek to highlight the computational advantages provided by block matrices. To this end, we compare the sketch compute times for block diagonal matrices with that of SRHT sketching matrices. We consider matrices $\mathbf{A}$ of sizes $2^{18} \times 40$, $2^{20} \times 40$ and $2^{22} \times 40$ and divide them into $J = 2^{10}$, 2^{12}, 2^{14} blocks respectively. In order to ensure fair comparison, we replace the SRHT matrix with randomly subsampled Fast Fourier transform (FFT) matrix, since both have the same theoretical sketch compute time, but the FFT matrix has very efficient software implementations. The sketch compute times are shown in Table 1. Our choice of J renders each block small enough for very efficient computations. This results in block diagonal matrices being much faster compared to the FFT matrix.

4 Proof Sketch

In this section, we provide a sketch of the proof for both Theorems 1 and 2. Full proofs are provided in the supplementary material. We first prove Theorem 1 and the proof for Theorem 2 follows by choosing $\mathbf{W}$ and $\mathbf{Y}$ appropriately, as explained in Section 2.2. The fundamental property of a distribution of matrices $\mathcal{D}$ that enables any $\mathbf{S} \sim \mathcal{D}$ to satisfy (8) is the subspace embedding moment property, defined in [Avron et al., 2016]:

$$\mathop{\mathrm{E}}_{\mathbf{S}\sim\mathcal{D}} \left\|(\mathbf{SU})^T(\mathbf{SU}) - \mathbf{I}\right\|^l \leq \epsilon^l\delta, \tag{12}$$

for some $l \geq 2$, where ϵ and δ are tolerance parameters that determine the sample complexity and $\mathbf{U}$ is any orthobasis for the span of the columns of $\mathbf{W}$ and $\mathbf{Y}$. Thus, our main goal is to prove the subspace embedding moment property holds for block diagonal sketching matrices.

Our methods differ from the common ϵ-net argument, since using union bound for block diagonal matrices results in a suboptimal sample complexity. The main tools we use are the estimates for the suprema of

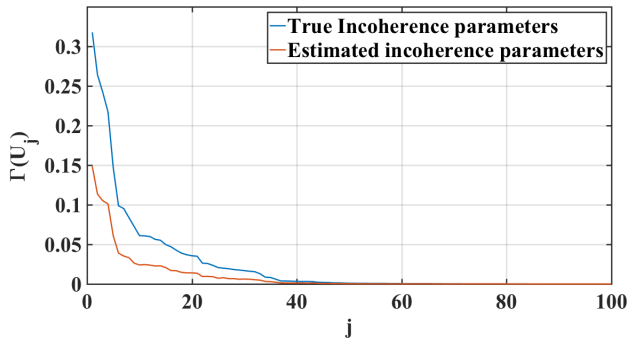

Figure 3: For a test matrix with $J = 100$, $\widetilde{N} = 10000$, the true incoherence values and the estimated values are within a constant factor of each other, shown here in a sorted. Choosing the block sizes M_j proportional to the estimated coherence parameters results in optimal sample complexities.

chaos processes found in [Krahmer et al., 2014] and an entropy estimate from the study of restricted isometry properties of block diagonal matrices computed in [Eftekhari et al., 2015]. We first establish tail bounds on the spectral norm of the matrix

$$\mathbf{\Delta} = (\mathbf{S}_D\mathbf{U})^T(\mathbf{S}_D\mathbf{U}) - \mathbf{I}, \tag{13}$$

where $\mathbf{U}$ is an orthobasis for a subspace of dimension d and then bound its moments to establish the subspace embedding moment property.

4.1 Tail bound on the spectral norm of the matrix $\mathbf{\Delta}$

We first express $\|\mathbf{\Delta}\|$ as

$$\begin{aligned}\|\mathbf{\Delta}\| &= \sup_{\substack{\mathbf{z}\in\mathbb{R}^d \\ \|\mathbf{z}\|=1}} \left|\mathbf{z}^T(\mathbf{S}_D\mathbf{U})^T(\mathbf{S}_D\mathbf{U})\mathbf{z} - 1\right| \\ &= \sup_{\substack{\mathbf{z}\in\mathbb{R}^d \\ \|\mathbf{z}\|=1}} \left|\|\mathbf{S}_D\mathbf{U}\mathbf{z}\|^2 - \mathrm{E}\,\|\mathbf{S}_D\mathbf{U}\mathbf{z}\|^2\right|. \end{aligned} \tag{14}$$

For the matrices $\mathbf{S}_j$, let $(\mathbf{S}_j)$ denote their vectorized versions, obtained by stacking the columns one below the other. Let $\mathbf{S}_v = [(\mathbf{S}_1)^T\ (\mathbf{S}_2)^T\ \cdots (\mathbf{S}_J)^T]^T$ be the vector containing all of the $(\mathbf{S}_j)$'s. Note that $\mathbf{S}_v$ is a vector with entries drawn from $\mathcal{N}(0,1)$. We can then express (14) as

$$\|\mathbf{\Delta}\| = \sup_{\mathbf{P}_z\in\mathcal{P}} \left|\|\mathbf{P}_z\mathbf{S}_v\|^2 - \mathrm{E}\,\|\mathbf{P}_z\mathbf{S}_v\|^2\right|$$

where $\mathcal{P}$ is defined as

$$\mathcal{P} = \left\{ \mathbf{P}_z = \begin{bmatrix} \mathbf{P}_1(z) & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2(z) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{P}_J(z) \end{bmatrix} \right\}$$

$$\mathbf{P}_j(z) = \frac{1}{\sqrt{M_j}} \begin{bmatrix} (U_1 z)^T & 0 & \cdots & 0 \\ 0 & (U_1 z)^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (U_1 z)^T \end{bmatrix}$$

where $z \in \mathbb{R}^d$ and $\|z\| = 1$. Observe that $\|\mathbf{\Delta}\|$ is then the supremum of the deviation of a Gaussian quadratic form from its expectation, taken over the set $\mathcal{P}$. This matches the framework developed in [Krahmer et al., 2014] to bound such suprema. We use their result (Theorem 3.1, [Krahmer et al., 2014]) to obtain tail bounds on $\|\mathbf{\Delta}\|$, stated in Lemma 1.

Lemma 1 *For any orthonormal matrix $\mathbf{U} \in R^{\widetilde{N}\times d}$ and a block diagonal matrix $\mathbf{S}_D$ as in Theorem 1, there exists a constant c such that*

$$\mathbb{P}\left(\|\mathbf{\Delta}\| \le c\sqrt{\frac{d\log(2/\delta)}{M_0}}\right) \ge 1-\delta. \tag{15}$$

For a desired tolerance ϵ, if $M_0 = \Omega\left(\frac{d\log(2/\delta)}{\epsilon^2}\right)$, $\mathbb{P}\left(\|\mathbf{\Delta}\| \le \epsilon\right) \ge 1-\delta$. This is similar to a subspace embedding guarantee. We now show that this tail bound naturally induces a bound on the moments of $\|\mathbf{\Delta}\|$, from which the main theorems in section 2 can be proved.

4.2 Moment bound on $\|\mathbf{\Delta}\|$

Tail bounds for certain random variables can be translated into bounds on their moments using the following result:

Lemma 2 (7.13, [Foucart and Rauhut, 2013]) *Suppose that a random variable q satisfies $\mathbb{P}\left(|q| \ge \mathrm{e}^{1/\gamma}\alpha u\right) \le \beta\mathrm{e}^{-u^\gamma/\gamma}$ for some $\gamma > 0$ and for all $u > 0$. Then, for $p > 0$, $\mathrm{E}\,|q|^p \le \beta\alpha^p(\mathrm{e}\gamma)^{p/\gamma}\Gamma\left(\frac{p}{\gamma}+1\right)$ where $\Gamma(\cdot)$ is the Gamma function.*

By choosing $q = \|\mathbf{\Delta}\|$, $\gamma = 2$, $\beta = 1$ and $\mathrm{e}^{-u^2/2} = \delta$, we obtain

Lemma 3 *For any orthonormal matrix $\mathbf{U} \in R^{\widetilde{N}\times d}$ and a block diagonal matrix $\mathbf{S}_D$ as in Theorem 1, if $M_0 = \Omega\left(\frac{d\log(2/\delta)}{\epsilon^2}\right)$, then for $p = (\frac{\log(1/\delta)}{\epsilon^2})$,*

$$\mathrm{E}\,\|\mathbf{\Delta}\|^p \le \epsilon^p\delta \tag{16}$$

Approximate matrix product guarantee Let

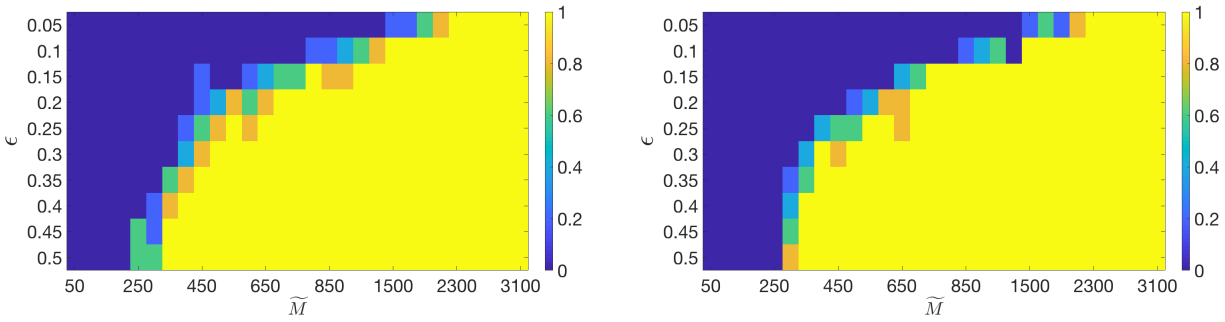

Figure 4: Each plot shows the empirical probability of $f(\hat{x}) \leq (1+\epsilon)f(x^*)$ for various values of $\widetilde{M}$, computed using an average over 10 trials. The left pane is for results with dense matrices with sub-Gaussian entries, the right pane for results with block diagonal sketching matrices.

Sketch compute time in seconds for large scale matrices				
$\widetilde{N},\ J$	$\widetilde{M} = 600$	$\widetilde{M} = 1400$	$\widetilde{M} = 2200$	$\widetilde{M} = 3000$
$2^{18}, 2^{10}$	$0.26; \mathbf{1.4 \times 10^{-2}}$	$0.26; \mathbf{2 \times 10^{-2}}$	$0.26; \mathbf{3.88 \cdot 10^{-2}}$	$0.26; \mathbf{4.2 \times 10^{-2}}$
$2^{20}, 2^{12}$	$1.16; \mathbf{2.7 \times 10^{-2}}$	$1.16; \mathbf{3.9 \times 10^{-2}}$	$1.16; \mathbf{5.1 \times 10^{-2}}$	$1.16; \mathbf{6.3 \times 10^{-2}}$
$2^{22}, 2^{14}$	$5.87; \mathbf{7.9 \times 10^{-2}}$	$5.87; \mathbf{9.1 \times 10^{-2}}$	$5.87; \mathbf{11 \times 10^{-2}}$	$5.86; \mathbf{11 \times 10^{-2}}$

Table 1: Sketch compute time in sec. for various matrix sizes $\widetilde{N}$ and sketch sizes $\widetilde{M}$. In each cell, the left figure for FFT sketch and the right figure in boldface is for block diagonal matrices.

$\mathbf{W}$ and $\mathbf{Y}$ be as in (8). As explained in [Cohen et al., 2015], we can assume that they have orthogonal columns. For a given k as in (8), let $\mathbf{W}$ and $\mathbf{Y}$ be partitioned into groups of k columns, with $\mathbf{W}_l$ and $\mathbf{Y}_{l'}$ denoting the l^{th} groups. The approach in [Cohen et al., 2015] then uses the following result in their argument, which follows from (16):

$$\mathrm{E}\left\|(\mathbf{S}\mathbf{W}_l)^T(\mathbf{S}\mathbf{Y}_{l'}) - \mathbf{W}_l^T\mathbf{Y}_{l'}\right\|^p \leq \epsilon^p \left\|\mathbf{W}_l\right\|^p \left\|\mathbf{Y}_{l'}\right\|^p \delta \tag{17}$$

for all pairs (l, l'). In their setting, this holds since the sketching is oblivious to the data matrices. Although block diagonal matrices are not oblivious, this result holds with for $M_0 = \Omega\left(\frac{2k\log(2/\delta)}{\epsilon^2}\right)$. This is because of the observation that if $\mathbf{U}$ is an orthobasis for the span of $\mathbf{W}$ and $\mathbf{Y}$ and $\mathbf{U}^{l,l'}$ is an orthobasis for the span of $\mathbf{W}_l$ and $\mathbf{Y}_{l'}$, then $\Gamma(\mathbf{U}_j^{l,l'}) \leq \Gamma(\mathbf{U}_j)$ for all pairs (l, l'). Hence, a given block diagonal sketching matrix $\mathbf{S}_D$ can satisfy (17). The rest of the proof remains the same as in [Cohen et al., 2015]. This concludes the proof for Theorem 1.

5 Conclusion

In this paper, we study a particular model that can be used while applying sketching techniques to high dimensional data that are available in a distributed fashion. Our proposed block diagonal sketching model forms an intermediate model between sampling methods and random projection methods and is a useful abstraction. We show theoretically and experimentally that choosing the sketch sizes proportional to a certain coherence term of the data blocks results in an optimal sample complexity. While we do not provide formal analysis of the algorithm to estimate the coherence parameters, we show empirically that they can be estimated.

6 Acknowledgements

This work was supported in part NSF CCF-1718771, NSF DMS 18-00872 and in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The authors would also like to acknowledge Agniva Chowdhury for their valuable discussion and feedback during the preparation of this paper.

References

[Ailon and Chazelle, 2006] Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM.

[Avron et al., 2017] Avron, H., Clarkson, K., and Woodruff, D. (2017). Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138.

[Avron et al., 2016] Avron, H., Clarkson, K. L., and Woodruff, D. P. (2016). Sharper bounds for regularized data fitting. *arXiv preprint arXiv:1611.03225*.

[Chen et al., 2015] Chen, S., Liu, Y., Lyu, M. R., King, I., and Zhang, S. (2015). Fast relative-error approximation algorithm for ridge regression. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pages 201–210, Arlington, Virginia, United States. AUAI Press.

[Chowdhury et al., 2018] Chowdhury, A., Yang, J., and Drineas, P. (2018). An iterative, sketching-based framework for ridge regression. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 989–998, Stockholmsmässan, Stockholm Sweden. PMLR.

[Clarkson and Woodruff, 2017] Clarkson, K. L. and Woodruff, D. P. (2017). Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54.

[Cohen et al., 2015] Cohen, M. B., Nelson, J., and Woodruff, D. P. (2015). Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*.

[Eftekhari et al., 2015] Eftekhari, A., Yap, H. L., Rozell, C. J., and Wakin, M. B. (2015). The restricted isometry property for random block diagonal matrices. *Applied and Computational Harmonic Analysis*, 38(1):1–31.

[Foucart and Rauhut, 2013] Foucart, S. and Rauhut, H. (2013). *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel.

[Krahmer et al., 2014] Krahmer, F., Mendelson, S., and Rauhut, H. (2014). Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904.

[McWilliams et al., 2014] McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G., and Vanchinathan, H. P. (2014). Loco: Distributing ridge regression with random projections. *stat*, 1050:26.

[Paul and Drineas, 2016] Paul, S. and Drineas, P. (2016). Feature selection for ridge regression with provable guarantees. *Neural Computation*, 28(4):716–742. PMID: 26890353.

[Wang et al., 2017] Wang, S., Gittens, A., and Mahoney, M. W. (2017). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *International Conference on Machine Learning*, pages 3608–3616.

[Woodruff, 2014] Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157.

[Yang et al., 2016] Yang, J., Meng, X., and Mahoney, M. W. (2016). Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, 104(1):58–92.