Sketching as a Tool for Numerical Linear Algebra

David P. Woodruff
IBM Research Almaden
dpwoodru@us.ibm.com

the essence of knowledge
Boston — Delft

Foundations and Trends® in Theoretical Computer Science

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

D. P. Woodruff. *Sketching as a Tool for Numerical Linear Algebra*. Foundations and Trends® in Theoretical Computer Science, vol. 10, no. 1-2, pp. 1–157, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-005-7

Foundations and Trends® in Theoretical Computer Science
Volume 10, Issue 1-2, 2014

Editorial Scope

Topics

Foundations and Trends® in Theoretical Computer Science publishes surveys and tutorials on the foundations of computer science. The scope of the series is broad. Articles in this series focus on mathematical approaches to topics revolving around the theme of efficiency in computing. The list of topics below is meant to illustrate some of the coverage, and is not intended to be an exhaustive list.

- Algorithmic game theory
- Computational algebra
- Computational aspects of combinatorics and graph theory
- Computational aspects of communication
- Computational biology
- Computational complexity
- Computational geometry
- Computational learning
- Computational Models and Complexity
- Computational Number Theory
- Cryptography and information security
- Data structures
- Database theory
- Design and analysis of algorithms
- Distributed computing
- Information retrieval
- Operations research
- Parallel algorithms
- Quantum computation
- Randomness in computation

Information for Librarians

Foundations and Trends® in Theoretical Computer Science, 2014, Volume 10, 4 issues. ISSN paper version 1551-305X. ISSN online version 1551-3068. Also available as a combined paper and online subscription.

Foundations and Trends® in
Theoretical Computer Science
Vol. 10, No. 1-2 (2014) 1–157

DOI: 10.1561/0400000060

Sketching as a Tool for Numerical Linear Algebra

David P. Woodruff
IBM Research Almaden
dpwoodru@us.ibm.com

Contents

Abstract

This survey highlights the recent advances in algorithms for numerical linear algebra that have come from the technique of linear sketching, whereby given a matrix, one first compresses it to a much smaller matrix by multiplying it by a (usually) random matrix with certain properties. Much of the expensive computation can then be performed on the smaller matrix, thereby accelerating the solution for the original problem. In this survey we consider least squares as well as robust regression problems, low rank approximation, and graph sparsification. We also discuss a number of variants of these problems. Finally, we discuss the limitations of sketching methods.

D. P. Woodruff. *Sketching as a Tool for Numerical Linear Algebra*. Foundations and Trends® in Theoretical Computer Science, vol. 10, no. 1-2, pp. 1–157, 2014.
DOI: 10.1561/0400000060.

1

Introduction

To give the reader a flavor of results in this survey, let us first consider the classical linear regression problem. In a special case of this problem one attempts to "fit" a line through a set of given points as best as possible.

For example, the familiar Ohm's law states that the voltage V is equal to the resistance R times the electrical current I, or $V = R \cdot I$. Suppose one is given a set of n example volate-current pairs (v_j, i_j) but does not know the underlying resistance. In this case one is attempting to find the unknown slope of a line through the origin which best fits these examples, where best fits can take on a variety of different meanings.

More formally, in the standard setting there is one *measured variable* b, in the above example this would be the voltage, and a set of d *predictor variables* $a_1, \ldots, a_d$. In the above example $d = 1$ and the single predictor variable is the electrical current. Further, it is assumed that the variables are linearly related up to a noise variable, that is $b = x_0 + a_1x_1 + \cdots + a_dx_d + \gamma$, where $x_0, x_1, \ldots, x_d$ are the coefficients of a hyperplane we are trying to learn (which does not go through the origin if $x_0 \neq 0$), and γ is a random variable which may be adversarially chosen, or may come from a distribution which we may have limited or no information about. The x_i are also known as the *model parameters*. By introducing an additional predictor variable a_0 which is fixed to 1, we can in fact assume that the unknown hyperplane goes through the origin, that is, it is an unknown subspace of codimension 1. We will thus assume that $b = a_1 x_1 + \cdots + a_d x_d + \gamma$ and ignore the affine component throughout.

In an experiment one is often given n observations, or n $(d+1)$-tuples $(a_{i,1}, \ldots, a_{i,d}, b_i)$, for $i = 1, 2, \ldots, n$. It is more convenient now to think of the problem in matrix form, where one is given an $n \times d$ matrix $\mathbf{A}$ whose rows are the values of the predictor variables in the d examples, together with an $n \times 1$ column vector $\mathbf{b}$ whose entries are the corresponding observations, and the goal is to output the coefficient vector $\mathbf{x}$ so that $\mathbf{Ax}$ and $\mathbf{b}$ are close in whatever the desired sense of closeness may mean. Notice that as one ranges over all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{Ax}$ ranges over all linear combinations of the d columns of $\mathbf{A}$, and therefore defines a d-dimensional subspace of $\mathbb{R}^n$, which we refer to as the column space of $\mathbf{A}$. Therefore the regression problem is equivalent to finding the vector $\mathbf{x}$ for which $\mathbf{Ax}$ is the closest point in the column space of $\mathbf{A}$ to the observation vector $\mathbf{b}$.

Much of the focus of this survey will be on the over-constrained case, in which the number n of examples is much larger than the number d of predictor variables. Note that in this case there are more constraints than unknowns, and there need not exist a solution $\mathbf{x}$ to the equation $\mathbf{Ax} = \mathbf{b}$.

Regarding the measure of fit, or closeness of $\mathbf{Ax}$ to $\mathbf{b}$, one of the most common is the least squares method, which seeks to find the closest point in Euclidean distance, i.e.,

$$\operatorname{argmin}_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2 = \sum_{i=1}^{n} (b_i - \langle \mathbf{A}_{i,*}, \mathbf{x} \rangle)^2,$$

where $\mathbf{A}_{i,*}$ denotes the i-th row of $\mathbf{A}$, and b_i the i-th entry of the vector $\mathbf{b}$. This error measure has a clear geometric interpretation, as the optimal $\mathbf{x}$ satisfies that $\mathbf{Ax}$ is the standard Euclidean projection of $\mathbf{b}$ onto the column space of $\mathbf{A}$. Because of this, it is possible to write the solution for this problem in a closed form. That is, necessarily one has $\mathbf{A}^T\mathbf{A}\mathbf{x}^* = \mathbf{A}^T\mathbf{b}$ for the optimal solution $\mathbf{x}^*$ by considering the gradient at a point $\mathbf{x}$, and observing that in order for it to be 0, that is for $\mathbf{x}$ to be a minimum, the above equation has to hold. The equation $\mathbf{A}^T\mathbf{A}\mathbf{x}^* = \mathbf{A}^T\mathbf{b}$ is known as the *normal equation*, which captures that the line connecting $\mathbf{A}\mathbf{x}^*$ to $\mathbf{b}$ should be perpendicular to the columns spanned by $\mathbf{A}$. If the columns of $\mathbf{A}$ are linearly independent, $\mathbf{A}^T\mathbf{A}$ is a full rank $d \times d$ matrix and the solution is therefore given by $\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. Otherwise, there are multiple solutions and a solution $\mathbf{x}^*$ of minimum Euclidean norm is given by $\mathbf{x}^* = \mathbf{A}^\dagger\mathbf{b}$, where $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$. Recall that if $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) of $\mathbf{A}$, where $\mathbf{U}$ is $n \times d$ with orthonormal columns, $\mathbf{\Sigma}$ is a diagonal $d \times d$ matrix with non-negative non-increasing diagonal entries, and $\mathbf{V}^T$ is a $d \times d$ matrix with orthonormal rows, then the Moore-Penrose pseudoinverse of $\mathbf{A}$ is the $d \times n$ matrix $\mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T$, where $\mathbf{\Sigma}^\dagger$ is a $d \times d$ diagonal matrix with $\mathbf{\Sigma}^\dagger_{i,i} = 1/\mathbf{\Sigma}_{i,i}$ if $\mathbf{\Sigma}_{i,i} > 0$, and is 0 otherwise.

The least squares measure of closeness, although popular, is somewhat arbitrary and there may be better choices depending on the application at hand. Another popular choice is the method of least absolute deviation, or ℓ_1-regression. Here the goal is to instead find $\mathbf{x}^*$ so as to minimize

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 = \sum_{i=1}^{n} |\mathbf{b}_i - \langle \mathbf{A}_{i,*}, \mathbf{x} \rangle|.$$

This measure is known to be less sensitive to outliers than the least squares measure. The reason for this is that one squares the value $\mathbf{b}_i - \langle \mathbf{A}_{i,*}, \mathbf{x} \rangle$ in the least squares cost function, while one only takes its absolute value in the least absolute deviation cost function. Thus, if $\mathbf{b}_i$ is significantly larger (or smaller) than $\langle \mathbf{A}_{i,*}, \mathbf{x} \rangle$ for the i-th observation, due, e.g., to large measurement noise on that observation, this requires the sought hyperplane $\mathbf{x}$ to be closer to the i-th observation when using the least squares cost function than when using the least absolute deviation cost function. While there is no closed-form solution for least absolute deviation regression, one can solve the problem up to machine precision in polynomial time by casting it as a linear programming problem and using a generic linear programming algorithm.

The problem with the above solutions is that on massive data sets, they are often too slow to be of practical value. Using näive matrix multiplication, solving the normal equations for least squares would take at least $n \cdot d^2$ time. For least absolute deviation regression, when casting the problem as a linear program one needs to introduce $O(n)$ variables (these are needed to enforce the absolute value constraints) and $O(n)$ constraints, and generic solvers would take $\text{poly}(n)$ time for an polynomial in n which is at least cubic. While these solutions are polynomial time, they are prohibitive for large values of n.

The starting point of this survey is a beautiful work by Tamás Sarlós [105] which observed that one could use *sketching techniques* to improve upon the above time complexities, if one is willing to settle for a randomized approximation algorithm. Here, one relaxes the problem to finding a vector $\mathbf{x}$ so that $\|\mathbf{Ax}-\mathbf{b}\|_p \leq (1+\varepsilon)\|\mathbf{Ax}^*-\mathbf{b}\|_p$, where $\mathbf{x}^*$ is the optimal hyperplane, with respect to the p-norm, for p either 1 or 2 as in the discussion above. Moreover, one allows the algorithm to fail with some small probability δ, which can be amplified by independent repetition and taking the best hyperplane found.

While sketching techniques will be described in great detail in the following chapters, we give a glimpse of what is to come below. Let $r \ll n$, and suppose one chooses a $r \times n$ random matrix $\mathbf{S}$ from a certain distribution on matrices to be specified. Consider the following algorithm for least squares regression:

1. Sample a random matrix $\mathbf{S}$.

2. Compute $\mathbf{S} \cdot \mathbf{A}$ and $\mathbf{S} \cdot \mathbf{b}$.

3. Output the exact solution x to the regression problem $\min_{\mathbf{x}} \|(\mathbf{SA})\mathbf{x} - (\mathbf{Sb})\|_2$.

Let us highlight some key features of this algorithm. First, notice that it is a *black box* reduction, in the sense that after computing $\mathbf{S} \cdot \mathbf{A}$ and $\mathbf{S} \cdot \mathbf{b}$, we then solve a smaller instance of least squares regression, replacing the original number n of observations with the smaller value of r. For r sufficiently small, we can then afford to carry out step 3, e.g., by computing and solving the normal equations as described above.

The most glaring omission from the above algorithm is which random familes of matrices $\mathbf{S}$ will make this procedure work, and for what values of r. Perhaps one of the simplest arguments is the following. Suppose $r = \Theta(d/\varepsilon^2)$ and $\mathbf{S}$ is a $r \times n$ matrix of i.i.d. normal random variables with mean zero and variance $1/r$, denoted $N(0, 1/r)$. Let $\mathbf{U}$ be an $n \times (d+1)$ matrix with orthonormal columns for which the column space of $\mathbf{U}$ is equal to the column space of $[\mathbf{A}, \mathbf{b}]$, that is, the space spanned by the columns of $\mathbf{A}$ together with the vector $\mathbf{b}$.

Consider the product $\mathbf{S}\cdot\mathbf{U}$. By 2-stability of the normal distribution, i.e., if $\mathbf{A} \sim N(0, \sigma_1^2)$ and $\mathbf{B} \sim N(0, \sigma_2^2)$, then $\mathbf{A} + \mathbf{B} \sim N(0, \sigma_1^2 + \sigma_2^2)$, each of the entries of $\mathbf{S} \cdot \mathbf{U}$ is distributed as $N(0, 1/r)$ (recall that the column norms of $\mathbf{U}$ are equal to 1). The entries in different rows of $\mathbf{S}\cdot\mathbf{U}$ are also independent since the rows of $\mathbf{S}$ are independent. The entries in a row are also independent by rotational invarance of the normal distribution, that is, if $\mathbf{g} \sim N(0, \mathbf{I}_n/r)$ is an n-dimensional vector of normal random variables and $\mathbf{U}_{*,1}, \ldots, \mathbf{U}_{*,d}$ are orthogonal vectors, then $\langle \mathbf{g}, \mathbf{U}_{*,1}\rangle, \langle \mathbf{g}, \mathbf{U}_{*,2}\rangle, \ldots, \langle \mathbf{g}, \mathbf{U}_{*,d+1}\rangle$ are independent. Here $\mathbf{I}_n$ is the $n\times n$ identity matrix (to see this, by rotational invariance, these $d+1$ random variables are equal in distribution to $\langle \mathbf{g}, \mathbf{e}_1\rangle, \langle \mathbf{g}, \mathbf{e}_2\rangle, \ldots, \langle \mathbf{g}, \mathbf{e}_{d+1}\rangle$, where $\mathbf{e}_1, \ldots, \mathbf{e}_{d+1}$ are the standard unit vectors, from which independence follows since the coordinates of g are independent).

It follows that $\mathbf{S}\cdot\mathbf{U}$ is an $r\times(d+1)$ matrix of i.i.d. $N(0, 1/r)$ random variables. For $r = \Theta(d/\varepsilon^2)$, it is well-known that with probability $1 - \exp(-d)$, all the singular values of $\mathbf{S} \cdot \mathbf{U}$ lie in the interval $[1-\varepsilon, 1+\varepsilon]$. This can be shown by arguing that for any fixed vector $\mathbf{x}$, $\|\mathbf{S} \cdot \mathbf{U}\mathbf{x}\|_2^2 = (1\pm\varepsilon)\|\mathbf{x}\|_2^2$ with probability $1-\exp(-d)$, since, by rotational invariance of the normal distribution, $\mathbf{S} \cdot \mathbf{U}\mathbf{x}$ is a vector of r i.i.d. $N(0, \|x\|_2^2)$ random variables, and so one can apply a tail bound for $\|\mathbf{S} \cdot \mathbf{U}\mathbf{x}\|_2^2$, which itself is a χ^2-random variable with r degrees of freedom. The fact that all singular values of $\mathbf{S}\cdot\mathbf{U}$ lie in $[1-\varepsilon, 1+\varepsilon]$ then follows by placing a sufficiently fine net on the unit sphere and applying a union bound to all net points; see, e.g., Theorem 2.1 of [104] for further details.

Hence, for all vectors $\mathbf{y}$, $\|\mathbf{S}\mathbf{U}\mathbf{y}\|_2 = (1\pm\varepsilon)\|\mathbf{U}\mathbf{y}\|_2$. But now consider the regression problem $\min_{\mathbf{x}} \|(\mathbf{S}\mathbf{A})\mathbf{x} - (\mathbf{S}\mathbf{b})\|_2 = \min_{\mathbf{x}} \|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2$. For each vector x, $\mathbf{A}\mathbf{x} - \mathbf{b}$ is in the column space of $\mathbf{U}$, and therefore by the previous paragraph, $\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$. It follows that by solving the regression problem $\min_x \|(\mathbf{S}\mathbf{A})\mathbf{x} - (\mathbf{S}\mathbf{b})\|_2$, we obtain a $(1 + \varepsilon)$-approximation to the original regression problem with probability $1 - \exp(-d)$.

The above technique of replacing $\mathbf{A}$ by $\mathbf{S} \cdot \mathbf{A}$ is known as a sketching technique and $\mathbf{S} \cdot \mathbf{A}$ is referred to as a (linear) sketch of $\mathbf{A}$. While the above is perhaps the simplest instantiation of sketching, notice that it does not in fact give us a faster solution to the least squares regression problem. This is because, while solving the regression problem $\min_{\mathbf{x}} \|(\mathbf{S}\mathbf{A})\mathbf{x} - (\mathbf{S}\mathbf{b})\|_2$ can now be done näively in only $O(rd^2)$ time, which no longer depends on the large dimension n, the problem is that $\mathbf{S}$ is a dense matrix and computing $\mathbf{S} \cdot \mathbf{A}$ may now be too slow, taking $\Theta(nrd)$ time.

Thus, the bottleneck in the above algorithm is the time for matrix-matrix multiplication. Tamás Sarlós observed [105] that one can in fact choose $\mathbf{S}$ to come from a much more structured random family of matrices, called fast Johnson-Lindenstrauss transforms [2]. These led to roughly $O(nd \log d) + \text{poly}(d/\varepsilon)$ time algorithms for the least squares regression problem. Recently, Clarkson and Woodruff [27] improved upon the time complexity of this algorithm to obtain *optimal* algorithms for approximate least squares regression, obtaining $O(\text{nnz}(\mathbf{A})) + \text{poly}(d/\varepsilon)$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries of the matrix $\mathbf{A}$. We call such algorithms input-sparsity algorithms, as they exploit the number of non-zero entries of $\mathbf{A}$. The $\text{poly}(d/\varepsilon)$ factors were subsequently optimized in a number of papers [92, 97, 18], leading to optimal algorithms even when $\text{nnz}(\mathbf{A})$ is not too much larger than d.

In parallel, work was done on reducing the dependence on ε in these algorithms from polynomial to polylogarithmic. This started with work of Rokhlin and Tygert [103] (see also the Blendenpik algorithm [8]), and combined with the recent input sparsity algorithms give a running time of $O(\text{nnz}(\mathbf{A}) \log(1/\varepsilon)) + \text{poly}(d)$ for least squares regression [27]. This is significant for high precision applications of least squares regression, for example, for solving an equation of the form $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Such equations frequently arise in interior point methods for linear programming, as well as iteratively reweighted least squares regression, which is a subroutine for many important problems, such as logistic regression; see [94] for a survey of such techniques for logistic regression. In these examples $\mathbf{A}$ is often formed from the Hessian of a Newton step in an iteration. It is clear that such an equation is just a regression problem in disguise (in the form of the normal equations), and the (exact) solution of $\operatorname{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ provides such a solution. By using high precision approximate regression one can speed up the iterations in such algorithms.

Besides least squares regression, related sketching techniques have also been instrumental in providing better robust ℓ_1-regression, low rank approximation, and graph sparsifiers, as well as a number of variants of these problems. We will cover these applications each in more detail.

Roadmap: In the next chapter we will discuss least squares regression in full detail, which includes applications to constrained and structured regression. In Chapter 3, we will then discuss ℓ_p-regression, including least absolute deviation regression. In Chapter 4 we will dicuss low rank approximation, while in Chapter 5, we will discuss graph sparsification. In Chapter 6, we will discuss the limitations of sketching techniques. In Chapter 7, we will conclude and briefly discuss a number of other directions in this area.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *ACM Symposium on Theory of Computing (STOC)*, 2006.

[3] Nir Ailon and Edo Liberty. Fast dimension reduction using rademacher series on dual bch codes. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008.

[4] Alexandr Andoni. High frequency moment via max stability. Available at http://web.mit.edu/andoni/www/papers/fkStable.pdf, 2012.

[5] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms via precision sampling. In *FOCS*, pages 363–372, 2011.

[6] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *APPROX-RANDOM*, pages 272–279, 2006.

[7] H. Auerbach. *On the Area of Convex Curves with Conjugate Diameters*. PhD thesis, University of Lwów, Lwów, Poland, 1930. (in Polish).

[8] Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM J. Scientific Computing*, 32(3):1217–1236, 2010.

[9] Haim Avron, Huy L. Nguyễn, and David P. Woodruff. Subspace embeddings for the polynomial kernel. In *NIPS*, 2014.

[10] Haim Avron, Vikas Sindhwani, and David P. Woodruff. Sketching structured matrices for faster nonlinear regression. In *NIPS*, pages 2994–3002, 2013.

[11] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. Comput. Syst. Sci.*, 74(1):97–114, 2008.

[12] Z. Bai and Y.Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability 21*, 3:1275–1294, 1993.

[13] Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David P. Woodruff. Fast and communication efficient algorithms for distributd pca. In *NIPS*, 2014.

[14] J.D. Batson, D.A. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 255–262. ACM, 2009.

[15] Michael W Berry, Shakhina A Pulatova, and GW Stewart. Algorithm 844: Computing sparse reduced-rank approximations to sparse matrices. *ACM Transactions on Mathematical Software (TOMS)*, 31(2):252–269, 2005.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

[17] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math*, 162:73–141, 1989.

[18] Jean Bourgain and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *CoRR*, abs/1311.2542, 2013.

[19] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near optimal column based matrix reconstruction. *SIAM Journal on Computing (SICOMP)*, 2013.

[20] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms (SODA)*, pages 968–977, 2009.

[21] Christos Boutsidis and David P. Woodruff. Optimal cur matrix decompositions. In *STOC*, pages 353–362, 2014.

[22] J.P. Brooks and J.H. Dulá. The ℓ_1-norm best-fit hyperplane problem. Technical report, Optimization Online, 2009.

[23] J.P. Brooks, J.H. Dulá, and E.L. Boone. A pure ℓ_1-norm principal component analysis. Technical report, Optimization Online, 2010.

[24] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[25] K. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.

[26] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. In *SODA*, pages 466–477, 2013.

[27] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *In STOC*, 2013.

[28] K.L. Clarkson and D.P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st annual ACM symposium on Theory of computing (STOC)*, 2009.

[29] Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. *Arxiv preprint arXiv:1410.6801*, 2014.

[30] Michael Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *Manuscript*, 2014.

[31] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. *Arxiv preprint arXiv:1408.5099*, 2014.

[32] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

[33] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson: Lindenstrauss transform. In *STOC*, pages 341–350, 2010.

[34] A. Deshpande and K. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 641–650. ACM, 2007.

[35] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(12):225–247, 2006.

[36] D. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 1989.

[37] P. Drineas and R. Kannan. Pass efficient algorithms for approximating large matrices. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 223–232, 2003.

[38] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal of Computing*, 36(1):184–206, 2006.

[39] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.

[40] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, volume 4168 of *Lecture Notes in Computer Science*, pages 304–314. Springer, 2006.

[41] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal Matrix Analysis and Applications*, 30(2):844–881, 2008.

[42] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.

[43] P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation, Technical Report, arXiv:0710.1435, 2007.

[44] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[45] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.

[46] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *ESA*, pages 304–314, 2006.

[47] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, 2005.

[48] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proc. 41th Annu. ACM Symp. on Theory of Computing (STOC), to appear*, 2011.

[49] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *SODA*, pages 630–649, 2010.

[50] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.

[51] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. In *STOC*, pages 475–484, 2010.

[52] Alex Gittens and Michael W Mahoney. Revisiting the nystrom method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.

[53] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.

[54] S.A. Goreinov, EE Tyrtyshnikov, and NL Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1-3):1–21, 1997.

[55] S.A. Goreinov, N.L. Zamarashkin, and E.E. Tyrtyshnikov. Pseudo-skeleton approximations by matrices of maximal volume. *Mathematical Notes*, 62(4):515–519, 1997.

[56] M. Gu and L. Miranian. Strong rank revealing Cholesky factorization. *Electronic Transactions on Numerical Analysis*, 17:76–92, 2004.

[57] Venkatesan Guruswami and Ali Kemal Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1207–1214. SIAM, 2012.

[58] Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

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

[60] Moritz Hardt and David P. Woodruff. How robust are linear sketches to adaptive inputs? In *STOC*, pages 121–130, 2013.

[61] T.M. Hwang, W.W. Lin, and D. Pierce. Improved bound for rank revealing LU factorizations. *Linear algebra and its applications*, 261(1-3):173–186, 1997.

[62] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[63] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[64] Piotr Indyk. Uncertainty principles, extractors, and explicit embeddings of l2 into l1. In *STOC*, pages 615–620, 2007.

[65] Yuri Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer, 1st edition, 2002.

[66] William Johnson and Gideon Schechtman. Very tight embeddings of subspaces of l_p, $1 = p < 2$, into ℓ_p^n. *Geometric and Functional Analysis*, 13(4):845–851, 2003.

[67] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *J. ACM*, 61(1):4, 2014.

[68] Ravi Kannan, Santosh Vempala, and David P. Woodruff. Principal component analysis and higher correlations for distributed data. In *COLT*, pages 1040–1057, 2014.

[69] Michael Kapralov, Yin Tat Lee, Cameron Musco, Christopher Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. *CoRR*, abs/1407.1289, 2014.

[70] Q. Ke and T. Kanade. Robust subspace computation using ℓ_1 norm, 2003. Technical Report CMU-CS-03-172, Carnegie Mellon University, Pittsburgh, PA.

[71] Qifa Ke and Takeo Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR (1)*, pages 739–746, 2005.

[72] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[73] Rafał Latała. Estimates of moments and tails of Gaussian chaoses. *Ann. Probab.*, 34(6):2315–2331, 2006.

[74] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[75] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 2001.

[76] Joseph Lehec. Moments of the Gaussian chaos. In *Séminaire de Probabilités XLIII*, volume 2006 of *Lecture Notes in Math.*, pages 327–340. Springer, Berlin, 2011.

[77] D. Lewis. Finite dimensional subspaces of l_p. *Studia Math*, 63:207–211, 1978.

[78] Mu Li, Gary L. Miller, and Richard Peng. Iterative row sampling. In *FOCS*, pages 127–136, 2013.

[79] Yi Li, Huy L. Nguyễn, and David P. Woodruff. On sketching matrix norms and the top singular vector. In *SODA*, 2014.

[80] D.G. Luenberger and Y. Ye. *Linear and nonlinear programming*, volume 116. Springer Verlag, 2008.

[81] M. Magdon-Ismail. Row Sampling for Matrix Algorithms via a Non-Commutative Bernstein Bound. *Arxiv preprint arXiv:1008.0587*, 2010.

[82] Malik Magdon-Ismail. Using a non-commutative bernstein bound to approximate some matrix algorithms in the spectral norm. *CoRR*, abs/1103.5453, 2011.

[83] A. Magen and A. Zouzias. Low Rank Matrix-valued Chernoff Bounds and Approximate Matrix Multiplication. *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2011.

[84] Avner Magen. Dimensionality reductions in l_2 that preserve volumes and distance to affine spaces. *Discrete & Computational Geometry*, 38(1):139–153, 2007.

[85] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[86] Michael W. Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences (PNAS)*, 106:697–702, 2009.

[87] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[88] O. L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1997.

[89] Jiri Matousek. *Lectures on Discrete Geometry*. Springer, 2002.

[90] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *Journal of Inequalities in Pure and Applied Mathematics*, 4(1:15), 2003.

[91] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A Parallel Iterative Solver for Strongly Over- or Under-Determined Systems. *ArXiv e-prints*, September 2011.

[92] Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *In STOC*, pages 91–100. ACM, 2013.

[93] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

[94] Thomas P. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Microsoft, 2003.

[95] L. Miranian and M. Gu. Strong rank revealing LU factorizations. *Linear Algebra and its Applications*, 367(C):1–16, July 2003.

[96] S. Muthukrishnan. *Data streams: algorithms and applications*. Foundations and Trends in Theoretical Computer Science, 2005.

[97] Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *In FOCS*, 2013.

[98] Jelani Nelson and Huy L. Nguyên. Lower bounds for oblivious subspace embeddings. In *ICALP (1)*, pages 883–894, 2014.

[99] Jelani Nelson and Huy L. Nguyễn. Sparsity lower bounds for dimensionality reducing maps. In *STOC*, pages 101–110, 2013.

[100] Huy Le Nguyễn. Personal communication, 2013.

[101] C.T. Pan. On the existence and computation of rank-revealing LU factorizations. *Linear Algebra and its Applications*, 316(1-3):199–222, 2000.

[102] Oded Regev. Personal communication, 2014.

[103] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212, 2008.

[104] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *CoRR*, abs/1003.2990v2, 2010.

[105] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

[106] Schechtman. More on embedding subspaces of l_p into ℓ_r^n. *Composition Math*, 61:159–170, 1987.

[107] N.D. Shyamalkumar and K. Varadarajan. Efficient subspace approximation algorithms. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 532–540, 2007.

[108] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *STOC*, pages 755–764, 2011.

[109] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *STOC*, pages 81–90, 2004.

[110] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011.

[111] N. Srivastava and D.A. Spielman. Graph sparsifications by effective resistances. In *Proceedings of the 40th ACM Symposium on Theory of Computing (STOC)*, 2008.

[112] G.W. Stewart. Four algorithms for the efficient computation of truncated QR approximations to a sparse matrix. *Numerische Mathematik*, 83:313–323, 1999.

[113] Michel Talagrand. Embedding subspaces of l_1 into ℓ_1^n. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990.

[114] Zhihui Tang. *Fast Transforms Based on Structured Matrices With Applications to The Fast Multipole Method*. PhD thesis, PhD Thesis, University of Maryland College Park, 2004.

[115] Mikkel Thorup and Yin Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 41(2):293–331, 2012.

[116] Joel Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images*, 2011.

[117] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 1st edition, 2008.

[118] E. Tyrtyshnikov. Mosaic-skeleton approximations. *Calcolo*, 33(1):47–57, 1996.

[119] E. Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton method. *Computing*, 64(4):367–380, 2000.

[120] Kasturi Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In *FSTTCS*, pages 486–497, 2012.

[121] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2011.

[122] S. Wang and Z. Zhang. Improving cur matrix decomposition and the nystrom approximation via adaptive sampling. *Journal of Machine Learning Research*, 2013.

[123] David P. Woodruff. Low rank approximation lower bounds in row-update streams. In *NIPS*, 2014.

[124] David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. *CoRR*, 2013.

[125] Dean Foster Yichao Lu, Paramveer Dhillon and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Proceedings of the Neural Information Processing Systems (NIPS) Conference*, 2013.

[126] Anastasios Zouzias. A matrix hyperbolic cosine algorithm and applications. In *ICALP (1)*, pages 846–858, 2012.