Asymptotic perspectives on sketching

Edgar Dobriban
based on joint works with Jonathan Lacotte, Sifan Liu, Mert Pilanci, David P. Woodruff, and Fan Yang

September 8, 2020

Overview

Overview

Overview

Sketched Linear Regression

Our Results

The Age of Data

- We live in the *Age of Data* (Statistics, Machine Learning, Artificial Intelligence, Data Science, ...)
- An enormous variety of digital data is generated and recorded every day
- Examples: Web (pages, links, ads, reviews), Science (health records, genetics, high-energy physics), ...

Big data

- It is expensive to analyze large datasets
  - Storage: Hard disk
  - Memory: RAM
  - Computation: CPU, GPU, TPU
  - Communication
- Fast algorithms are important, because they end up saving real resources (money, time, energy)

Big data

- Many approaches to efficiently processing big data (from computer science, statistics, optimization, machine learning, data mining, signal processing and information theory, etc)
  - Efficient data structures
  - Distributed and parallel computing
  - Data reduction and compression
  - Online and streaming algorithms
  - Randomized algorithms
  - ....

Randomized algorithms for processing matrices and data

- Many of data problems can be modeled via matrices
- E.g., Collect data on n patients. Measure their features (demographics, lifestyle, genetics, preferences, medical tests, health outcomes)
- Arrange features of patient i into $p \times 1$ vector x_i, and let $x_i^\top$ be the rows of data matrix $X \in \mathbb{R}^{n \times p}$.
- Data matrix X is big $\to$ reduce it to smaller matrix
- Sample size reduction: $S \in \mathbb{R}^{r \times n}$ $(r < n)$

$$X \in \mathbb{R}^{n \times p} \to SX \in \mathbb{R}^{r \times p}$$

- Dimension reduction: $T \in \mathbb{R}^{p \times t}$ $(p < t)$

$$X \in \mathbb{R}^{n \times p} \to XT \in \mathbb{R}^{n \times t}$$

- How to choose S, T? Randomize, leading to random projections

Uses of Random Projections

- Randomized numerical linear algebra [Drineas and Mahoney, 2016]: matrix multiplication, SVD, low rank approximation, etc.
- Statistics, Machine learning and Data mining: nonparametric regression [Yang et al., 2017], ridge regression [Lu et al., 2013], two sample testing [Lopes et al., 2011], classification [Cannings and Samworth, 2017], clustering [Fern and Brodley, 2003], PCA [Rokhlin et al., 2009], deep learning [Abadi et al., 2016], subspace clustering [Pimentel-Alarcón et al., 2016], tensor regression [Zhang et al., 2019], etc.
- Convex optimization [Pilanci and Wainwright, 2015, 2016, 2017]
- Econometrics [Ng, 2017]
- Genomics [Galinsky et al., 2016]

Comments

- RP/randomized algorithms:
- Strengths: General-purpose, works for arbitrary/adversarial data, fast,
- Limitations: May fail with some probability, results depend on random draw (large variance), hard to tune, unknown/mysterious when and how well they work
- We discuss our works on developing a "big data" asymptotic perspective on sketching.
- We shed light on the mystery, giving concrete tools for practitioners to be able to decide how and when to use random projections - "instruction manual" for RP/sketching
- Clean and elegant mathematical results

Our works on asymptotic perspectives on sketching

- Linear regression: [Dobriban and Liu, 2019], NeurIPS 2019
- Ridge regression: [Liu and Dobriban, 2019], ICLR 2020
- Iterative Hadamard Hessian Sketch: [Lacotte et al., 2020]
- PCA: [Yang et al., 2020]

Overview

Linear Regression

- Data:
  1. n samples/datapoints. Arrange features of sample i into $p \times 1$ vector x_i, and let $x_i^\top$ be the rows of data matrix $X \in \mathbb{R}^{n \times p}$, $n > p$.
  2. Outcomes y_i arranged into $n \times 1$ vector Y.
- Goal: understand effect of features on outcome
- Linear regression model: $Y = X\beta + \varepsilon$
  1. Unknown regression coefficients $\beta \in \mathbb{R}^p$,
  2. Noise $\varepsilon \in \mathbb{R}^n$. The ε_i's are uncorrelated, with $\mathbb{E}\left[\varepsilon_i\right] = 0$, $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$
- Ordinary Least Squares (OLS) estimator

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Sketched Linear Regression

- Complexity of full OLS: $O(np^2)$ floating point operations (flops).
- n, p are large, so this is too expensive.
- Approximation is necessary
- Sketch and solve: apply random projection/sketching matrix $S \in \mathbb{R}^{r \times n}$, then do least-squares on $(\tilde{X}, \tilde{Y}) = (SX, SY)$, which gives

$$\hat{\beta}_s = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}.$$

Questions

- Faster but less accurate... By how much?
- How does this work from a statistical point of view? What is the loss? How many dimensions do we need?
- How to choose sketching matrix S? for instance
  - Uniform sampling of rows of X (i.e., subsampling)
  - iid entries (Gaussian or $-1, 0, 1$),
  - Haar
  - randomized Hadamard/Fourier transform
  - leverage sampling
- Under what conditions on the data X?

Statistical efficiency

- How to measure the statistical efficiency?
- Standard measures for evaluating estimator $\hat{\beta}$ of parameter β
  - Mean Squared Error: $\mathbb{E}\left[\|\hat{\beta} - \beta\|^2\right]$
  - In regression context, Residual Error: $\mathbb{E}\left[\|Y - X\hat{\beta}\|^2\right]$
  - In regression context, Test Error: Let x_t, y_t be a test datapoint. We use $\hat{y}_t = x_t^\top \hat{\beta}$ to predict the unobserved y_t: $\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]$
- Approach inspired by [Raskutti and Mahoney, 2016]

Relative Efficiencies

- How to compare the statistical efficiency of two estimators?
- Relative efficiency: ratio of errors

$$\text{Variance efficiency: } VE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[\|\beta - \hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|\beta - \hat{\beta}\|^2\right]},$$

$$\text{Residual efficiency: } RE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[\|Y - X\hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|Y - X\hat{\beta}\|^2\right]},$$

$$\text{The out-of-sample efficiency: } OE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta}_s)^2\right]}{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]}.$$

Sketching Methods

- Three broad categories
  1. Sampling methods: sample rows of X independently, either iid or according to "importance" (fast/non-robust)
  2. iid entries: entries of S are iid (popular/easy to understand/less accurate/slow)
  3. Structured orthogonal: S consists of random rows from a structured orthonormal/unitary matrix such as Fourier/Hadamard transform (fast/accurate/robust/hard to analyze and understand)

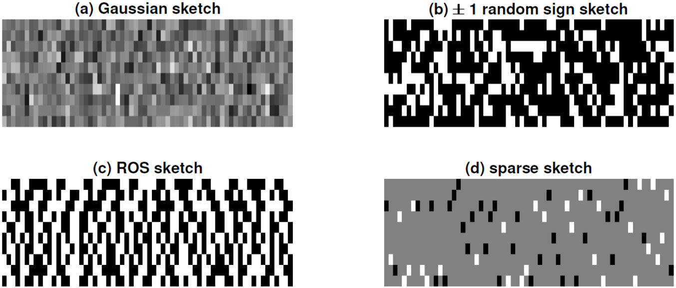


Figure: From review by Pilanci.

Sampling Methods

- Uniform sampling: S samples each row of X with equal probability
- Leverage score sampling: S samples each row of X w.r.t. its leverage scores $H_{ii} = x_i^\top (X^\top X)^{-1} x_i$, where $H = X(X^\top X)^{-1} X^\top$ is the "hat" matrix.

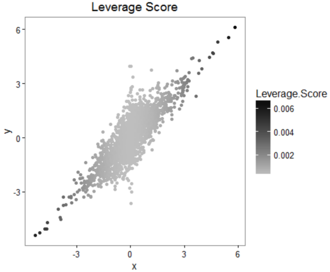


Figure: [Zhang et al., 2018] "Statistical leveraging methods in big data"

- Performing sampling: $O(n)$ flops
- Computing leverage scores: $O(np^2)$
- Estimating leverage scores by Hadamard transform: $O(pn \log n)$ [Drineas et al., 2012] (may use Hadamard transform directly)

iid entries

- Gaussian projection: S has iid $\mathcal{N}(0, 1)$ entries
- Sparse projection: S has iid random entries with $\mathbb{P}(S_{ij} = \pm 1) = 1/(2s)$, $\mathbb{P}(S_{ij} = 0) = 1 - 1/s$, where s is some fixed constant greater than 1.
- Generally: S_{ij} are iid standardized random variables
- Computing the projection SX: $O(rnp)$

Structured orthogonal

- Fourier transform:
  - Start with Discrete Fourier Transform matrix F, with entries $F_{uv} = n^{-1/2} e^{-2\pi i \frac{(u-1)(v-1)}{n}}$.
  - Let $S = DFQP$, where Q has iid ± 1 diagonals, D selects random rows, P is uniform permutation matrix
- Hadamard transform: Same construction, starting with recursively defined Hadamard matrix, $H_1 = 1$,

$$H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix},$$

- Computing SX via Fast Fourier transform: $O(pn \log n)$

Overview

Theoretical results

Table: Approximate efficiency of sketching. Original linear model: $n \times p$, with n samples and p dimensions ($n > p$). Sketched linear model: $r \times p$ ($n > r > p$). The loss functions are VE (variance efficiency) and OE (out-of-sample prediction efficiency).

<table>
<tr><th>X</th><th>S</th><th>$\mathrm{VE} = \frac{\mathbb{E}\left[\|\beta - \hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|\beta - \hat{\beta}\|^2\right]}$</th><th>$\mathrm{OE} = \frac{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta}_s)^2\right]}{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]}$</th></tr>
<tr><td rowspan="2">Arbitrary</td><td>iid entries</td><td>$1 + \frac{n-p}{r-p}$</td><td>$\frac{nr-p^2}{n(r-p)}$</td></tr>
<tr><td>Haar/Hadamard</td><td rowspan="2">$\frac{n-p}{r-p}$</td><td rowspan="2">$\frac{r(n-p)}{n(r-p)}$</td></tr>
<tr><td>Ortho-invariant</td><td>Uniform sampling</td></tr>
<tr><td>Elliptical: $WZ\Sigma^{\frac{1}{2}}$</td><td>Leverage sampling</td><td>$\frac{\eta_{sw^2}^{-1}(1-p/n)}{\eta_{w^2}^{-1}(1-p/n)}$</td><td>$\frac{1+\mathbb{E}\left[w^2\right]\eta_{sw^2}^{-1}(1-\gamma)}{1+\mathbb{E}[w^2]\eta_{w^2}^{-1}(1-\gamma)}$</td></tr>
</table>

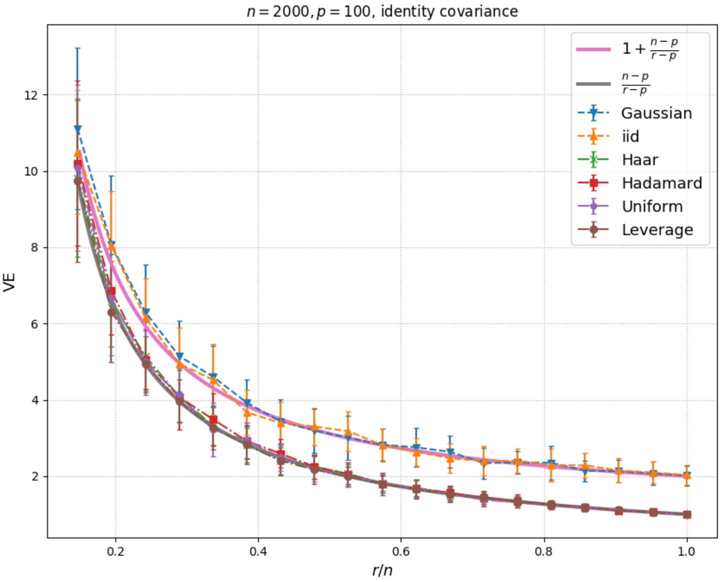
Comments on Our Results

- Accurate in numerical simulations and data examples
- Can be used as "sample size calculations" (do not depend on unknown quantities)
- Show separation between sketching methods: orthogonal sketches are better than iid sketching
- Hadamard was empirically observed to be accurate, but there was no proof it is better than iid projections (proof uses non-elementary random matrix theory)

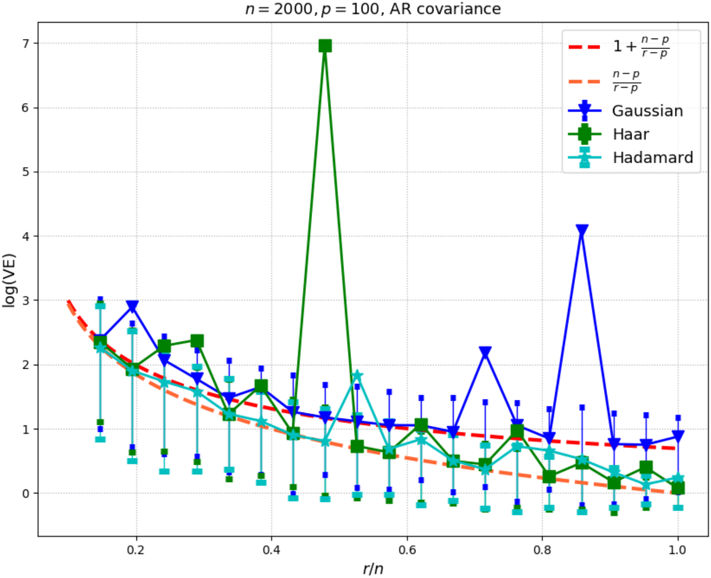
Numerical Experiments

- Take $n = 2000, p = 100$, with r ranging from 300 to 2000.
- Generate and fix the data matrix X from
  - a standard Gaussian distribution
  - multivariate t distribution with AR-1 covariance
- Generate β, ε as standard Gaussian
- At each dimension r, randomly generate 50 sketching matrices S
- Compute $\hat{\beta}_s$ and $\hat{\beta}$, then find VE

Numerical Results



Numerical Results



Empirical Data Analysis Results

- We make empirically testable predictions - how much do residuals increase after sketching?
- These can be tested using data without making any assumptions (very different from most work in high-dimensional statistics...)
- We test our results on the Million Song Year Prediction Dataset (MSD) [Bertin-Mahieux et al., 2011].
- $n = 515,344$ samples and $p = 90$ features
- We take a random test set of size 10,000.
- For each target dimension r, we show the mean, 5% and 95% quantiles over 10 repetitions.

Million Song Year Prediction Dataset

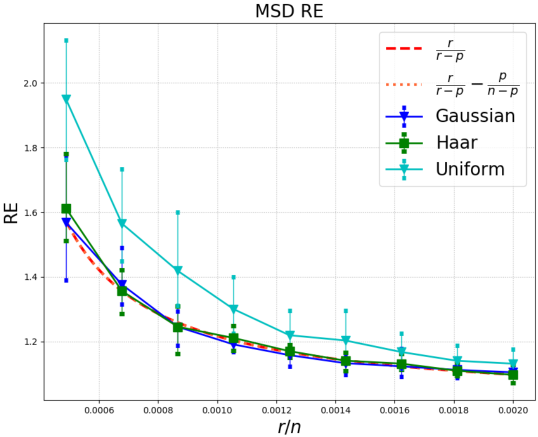


Figure: Good agreement of theory and empirical results for Gaussian and Hadamard/Haar projections. For uniform sampling, our theory requires the data matrix X to be rotationally invariant, which may not hold, leading to less accuracy.

Proof aspects

- Our proofs rely on asymptotic random matrix theory and free probability
- Perfect fit: algorithm random, need weak assumptions
- "Standard" results (such as the Marchenko-Pastur law) are *not* enough.
- To study the subsampled randomized Hadamard transform (SRHT), we discovered that we can use the results of [Anderson and Farrell, 2014], on *asymptotically liberating sequences.* (see also Tulino, Caire, Shamai, Verdu 2010)

Anderson & Farrell's paper

Asymptotically liberating sequences of random unitary matrices

Greg W. Anderson [a,*], Brendan Farrell [b,1]

[a] *School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA*
[b] *Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA*

A R T I C L E   I N F O

Article history:
Received 22 February 2013
Accepted 21 December 2013
Available online 29 January 2014
Communicated by Dan Voiculescu

MSC:
60B20
42A61
46L54
15B52

Keywords:
Free probability

A B S T R A C T

A fundamental result of free probability theory due to Voiculescu and subsequently refined by many authors states that conjugation by independent Haar-distributed random unitary matrices delivers asymptotic freeness. In this paper we exhibit many other systems of random unitary matrices that, when used for conjugation, lead to freeness. We do so by first proving a general result asserting "asymptotic liberation" under quite mild conditions, and then we explain how to specialize these general results in a striking way by exploiting Hadamard matrices. In particular, we recover and generalize results of the second-named author and of Tulino, Caire, Shamai and Verdú.

Proof outline for VE

- Expand $\mathbb{E}\|\beta - \hat{\beta}_s\|^2$ in terms of trace:

$$\mathbb{E}\|\beta - \hat{\beta}_s\|^2 = \mathbb{E}\,\mathrm{tr}[QQ^\top] = \|Q\|_{Fr}^2$$
$$Q = (X^\top S^\top S X)^{-1} X^\top S^\top S$$

- Calculate
  - Finite-sample expectation for Gaussian S (using *Wishart properties*)
  - Limiting expectation for iid S (using *Lindeberg swapping* [Chatterjee, 2006])
  - For orthogonal S, reduces to $\mathrm{tr}(X^\top S^\top S X)^{-1}$

Proof outline for VE, Hadamard/Fourier

- Find limit of $\mathrm{tr}(X^\top S^\top SX)^{-1}$, X arbitrary, $S = DHQP$
  - $D = diag(B_i)$, $B_i \sim Bernoulli(r/n)$ selects random rows,
  - H is Hadamard,
  - Q has iid ± 1 diagonals,
  - P is uniform permutation matrix
- Let $W = P^\top QHQP$ be the *signed Hadamard matrix*. Then

$$X^\top S^\top SX = X^\top (P^\top QH) D (HQP) X \overset{d}{=} X^\top WDWX.$$

- Anderson & Farrell: D, $M = n^{-1} WXX^\top W$ are asymptotically freely independent

What is asymptotic free independence?

- Asymptotic free independence is a version of independence for non-commutative random variables, such as matrices (Voiculescu, 1986)
- Intuitively: Two large-dimensional random matrices are freely independent, if their eigenspaces are completely decorrelated
- e.g., $X^T X$ where X has iid Gaussian entries, and any fixed matrix M
- Formal definition: sequences of $n \times n$ symmetric random matrices A_n, B_n a.s. freely indept if

$$n^{-1}\,\mathrm{tr}[(P_1(A_n) - \mathbb{E}P_1(A_n)) \cdot (Q_1(B_n) - \mathbb{E}Q_1(B_n)) \ldots] \to_{a.s.} 0$$

For any sequence of polynomials P_i, Q_i
(analog of $\mathbb{E}P(A)Q(B) = \mathbb{E}P(A)\mathbb{E}Q(B)$)

How do we use free independence?

- Allows us to decouple the traces
- Recall that we need to calculate $\operatorname{tr}(MD)^{-1}$, and M, D are asymptotically freely independent.
- The spectrum of matrices L is characterized through their Stieltjes transform $m_{L,n}(z) = n^{-1}\operatorname{tr}(L-z)^{-1}$.
- We know that the Stieltjes transforms of M, D converge to m_M, m_D; and free independence gives an equation for m_{MD} (in terms of S-transform)
- Find and solve that equation

Related Work

[Raskutti and Mahoney, 2016] showed that for Gaussian projection,

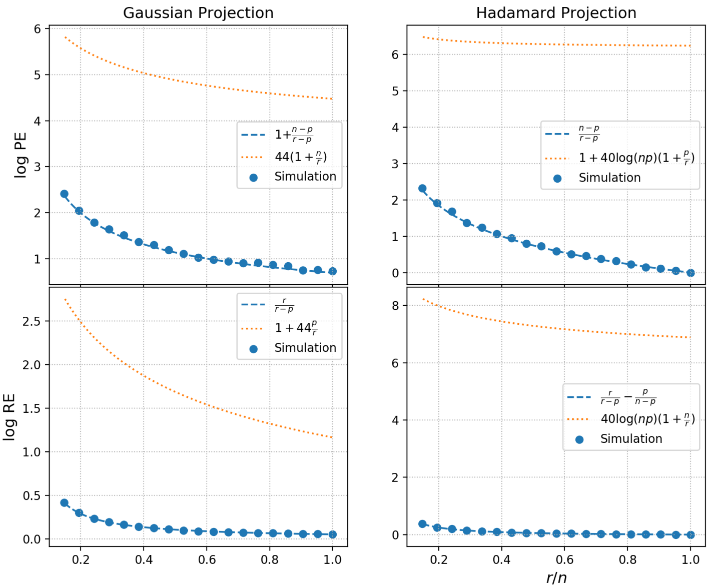
$$PE \leq 44(1 + \frac{n}{r}),$$
$$RE \leq 1 + 44\frac{p}{r},$$

and for Hadamard/Fourier transform,

$$PE \leq 1 + 40\log(np)(1 + \frac{p}{r}),$$
$$RE \leq 40\log(np)(1 + \frac{n}{r}),$$

hold with some constant probability.
Our contribution: We find the asymptotically exact answer.

Comparing results



Summary

- Asymptotic perspective on sketch-and-solve linear regression
    1. Similar approach for ridge regression, iterative Hessian sketch, PCA

Thanks!

References I

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

Greg W Anderson and Brendan Farrell. Asymptotically liberating sequences of random unitary matrices. *Advances in Mathematics*, 255:381–413, 2014.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

Sourav Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

References II

Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems*, pages 3675–3685, 2019.

Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

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(Dec):3475–3506, 2012.

Xiaoli Z Fern and Carla E Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 186–193, 2003.

Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

References III

Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci. Limiting spectrum of randomized hadamard transform and optimal iterative sketching methods. *arXiv preprint arXiv:2002.00864*, 2020.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2019.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.

Serena Ng. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, National Bureau of Economic Research, 2017.

References IV

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Daniel Pimentel-Alarcón, Laura Balzano, and Robert Nowak. Necessary and sufficient conditions for sketched subspace clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1335–1343. IEEE, 2016.

Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

References V

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

Fan Yang, Sifan Liu, Edgar Dobriban, and David P Woodruff. How to reduce dimension with pca and random projections? *arXiv preprint arXiv:2005.00511*, 2020.

Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

Anru Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. Islet: Fast and optimal low-rank tensor regression via importance sketching. *arXiv preprint arXiv:1911.03804*, 2019.

Xinlian Zhang, Rui Xie, and Ping Ma. Statistical leveraging methods in big data. In *Handbook of Big Data Analytics*, pages 51–74. Springer, 2018.