# RESEARCH STATEMENT

ELIZAVETA REBROVA

**My research is in high-dimensional probability, random matrix theory and mathematical data science.** In the modern world, we have access to a huge and only growing amount of data. However, to convert this rich data to actionable knowledge, we are confronted with the challenge of treating complex, frequently high-dimensional and large-scale data in a deliberate manner: a manner that captures large and important underlying trends, but does not miss small ones, that is practically efficient and scientifically justified. There are many ways in which a probabilistic view can help with this task, some standard examples include modeling a real-world dataset by an artificial one coming from a suitable distribution, or creating a randomized algorithm for an otherwise NP-hard problem that is fast and almost always right. This motivates the main theme of my research: **I study the structure of large high-dimensional objects in the presence of randomness, and use this understanding to develop randomized algorithms that efficiently process complex data.** The tools I use come from probability, functional analysis, convex geometry, optimization, numerical linear algebra, and machine learning. The objects of my study range from matrices, graphs and tensors to any instances of large data, and this places my work at the intersection of pure and applied mathematics. On the pure math side, my research uncovers the beauty of a special order that naturally appears *with high probability* in large random systems, such as, in the spectrum of large random matrices. On the applied side, my research makes steps to close the gap between slow evolving theory and ad-hoc industry practices. I develop data processing techniques that are simultaneously efficient, supported by theory, and well-interpretable. As an illustration of these principles, I will henceforth focus on three concrete areas of my work: (1) non-asymptotic **random matrix theory** in the heavy-tailed regime; (2) **tensor factorization and dimension reduction**, and (3) **randomized iterative methods** for solving linear systems.

## 1. Matrices: Extremal singular values of heavy-tailed random matrices.

A classical way to understand the structure of a random matrix $A$ is to look at its spectrum [Tao12, Ver16]. For example, the largest and the smallest singular values ($\sigma_k(A) := \sqrt{eig_k(A^TA)}$) determine the basic geometric properties of $A$ as a linear operator, namely, the norm of $A$ and its inverse:

$$\sigma_{max}(A) = \sup_{\|x\|_2=1} \|Ax\|_2 = \|A\|, \quad \sigma_{min}(A) = \inf_{\|x\|_2=1} \|Ax\|_2 = 1/\|A^{-1}\|.$$

Furthermore, the condition number of a matrix $\kappa(A) := \sigma_{max}(A)/\sigma_{min}(A)$ estimates its stability properties and serves as a crucial parameter in the algorithms analysis (e.g., [Dem97]). Note that a well-bounded condition number is a result of quantitatively good invertibility of the matrix (that is, $\|A^{-1}\|$ is not too large) and well-bounded operator norm (that is, $\|A\|$ is also not too large). For an $n \times n$ random matrix $A$ with independent standard Gaussian entries, both $\|A^{-1}\|$ and $\|A\|$ are of order $O(\sqrt{n})$ with high probability [Ede88]. The same upper bounds hold if we relax the distribution assumptions to any subgaussian distribution, that is, a distribution with tails that decay at least as fast as Gaussian distribution tails [RV08]. In our work with K. Tikhomirov [RT18], we establish that the upper bound on $\|A^{-1}\|$ holds for a significantly broader class of matrices. Our result contains the result of [RV08] as a special case, while our probabilistic bound is sharp, unlike those obtained in related work [TV08, TV10]:

**Theorem 1.** *Let $A$ be an $n \times n$ matrix for $n$ large enough with independent and identically distributed (i.i.d.) centered elements with unit variance (so-called* heavy-tailed matrix model*). Then for any $\varepsilon > 0$*

$$\mathbb{P}\big\{\|A^{-1}\| \geq \varepsilon^{-1}\sqrt{n}\big\} \leq L\varepsilon + u^n, \text{ with constants } L > 1 \text{ and } u \in (0,1).$$

However, for the heavy-tailed matrix model, $\|A\|$ might be much bigger (in particular, weak fourth moment is necessary for the convergence in probability of $\|A\|/\sqrt{n}$ when $n$ grows to infinity; see [Sil89]).

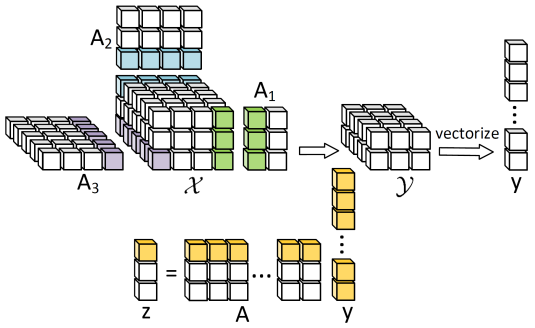


FIGURE 1. An example of 2-stage JL embedding applied to a 3-dimensional tensor $\mathcal{X} \in \mathbb{R}^{3\times4\times5}$. Next, the resulting tensor is vectorized to $\mathbf{y} \in \mathbb{R}^{24}$, and a $2^{\text{nd}}$-stage JL is then performed to obtain $\mathbf{z} = \mathbf{A}\mathbf{y}$ where $\mathbf{A} \in \mathbb{R}^{3\times24}$, and $\mathbf{z} \in \mathbb{R}^3$.

In the follow-up work with R.Vershynin [RV18], we show that the finite second moment condition is necessary and sufficient for the existence of a local regularization of $A$ that returns $\|A\|$ to its "ideal order" (see [RV18, Theorems 1.1, 1.3]) for the formal statements):

**Theorem 2** (Informal statement)**.** *Let $A$ be an $n \times n$ matrix with i.i.d. centered elements $a_{ij}$,*

- *if $a_{ij}$ have finite variance, then for every small $\varepsilon > 0$ of our choice, with high probability there exists an $\varepsilon n \times \varepsilon n$ submatrix $A_0$ so that if we zero out all the elements inside $A_0$,*
$$\|A \setminus A_0\| \le C_\varepsilon \sqrt{n}, \qquad \text{where } C_\varepsilon \sim \ln \varepsilon^{-1}/\sqrt{\varepsilon},$$
- *otherwise we have to zero out almost all of $A$ to bring the norm to the order $\sqrt{n}$.*

This is an existence result, which does not explain a way to find the small submatrix to be zeroed out. In my follow-up paper [Reb19], I present two constructive versions of the regularization procedure that achieve the $O(\sqrt{n \log\log n})$ order of the norm for heavy-tailed matrices with i.i.d. symmetrically distributed entries. A simple way is to remove an $\varepsilon$-fraction of the rows and columns with the largest norms, while a more sophisticated algorithm achieves the same goal by removing only a small submatrix as promised by Theorem 2. See more details in [Reb19, Theorem 1, Algorithm 1].

**Outlook:** It is very important to get closer to the complex distributions of real data and thus to get the results *beyond i.i.d. matrix models*. Here, even small generalizations often require development of new toolkits of math methods. Some particular examples of the theoretical problems I am interested in are (a) to estimate an upper bound on $\sigma_{min}(A+M)$, where $A$ is subgaussian and $M$ is an arbitrary non-random shift (b) to get a sharp probability estimate for the quantitative invertibility of rectangular matrices with entries distributed with bounded densities. The latter is needed, in particular, to provide convergence guarantees for the iterative randomized linear solvers (see Section 3 below) with non-Gaussian sketches.

Another very natural non-i.i.d. random matrix model is adjacency and Laplacian matrices associated with random graphs. My ongoing work with P. Salanevich is related to *signal processing on graphs*. One of the main results in classical signal processing is the uncertainty principle, stating that a signal cannot be simultaneously localized in time and frequency. For signals defined on graphs, uncertainty quantification is closely linked with the *delocalization properties of eigenvectors* – one of the most active topics in the modern random matrix theory. It quantifies the similarity between the matrix of eigenvectors $V$ and a standard Gaussian matrix (for example, in terms of the largest element of $V$ [BKY17] or the number of small elements in each column of $V$ [RV16, BL13]). We generalized the results obtained in [TV17] to obtain a better uncertainty bound in terms of the sub-blocks of $V$ and employed it to get an uncertainty principle for d-regular graphs using several known delocalization results. The next steps include establishing a new class of delocalization results addressing the relative sizes of eigenvectors (sub-blocks of $V$) and considering more general graph models.

## 2. Tensors: Modewise dimension reduction methods for tensors

Tensors (multi-way arrays), despite being direct higher-order generalizations of matrices, present many interesting mathematical non-trivialities. For example, the notion of the spectrum is not well-defined in

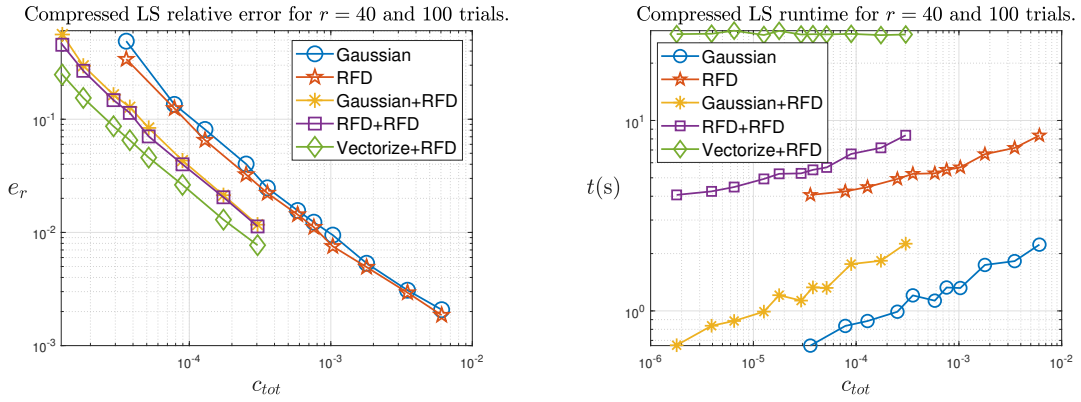


FIGURE 2. Modewise dimension reduction methods for tensor fitting achieve very similar relative error for the same total compression rate (left), but much more efficient (right)

the tensor case, and there are multiple ways to define the tensor rank. One of the most natural definitions is the so-called CP-rank: for a tensor $\mathcal{X}$, it is a minimal number of rank-one tensors (being outer products of a collection of vectors) whose linear combination constitutes $\mathcal{X}$, namely,

$$\mathcal{X} = \sum_{j=1}^{r} x_j^1 \otimes \ldots \otimes x_j^d \quad \text{for the } d\text{-way tensor of rank } r \text{ in } \mathbb{C}^{n_1 \times \cdots \times n_d}.$$

There are no efficient algorithms for computing the CP decomposition precisely. In fact, this problem has been proven to be NP-hard ([Hås90]). Moreover, many real-life tensors are only approximately low-rank due to noise, imperfect measurements, etc. Thus, it is a very important problem to approximate a given tensor by a low rank tensor (so-called *fitting problem*) in some norm (e.g., $\|\mathcal{X}\|^2 = \sum \mathcal{X}_{i_1 \ldots i_d}^2$).

In my work with M. Iwen, D. Needell and A. Zare [INRZ19], we propose and analyze *modewise oblivious* dimension reduction methods that speed up and reduce memory when solving the fitting problem. A common way to tackle the fitting problem (CPD-ALS, [KB09]) is to start with a randomly generated tensor, and then optimize its components mode by mode, iteratively finding the best fitting tensor in a fixed low-dimensional subspace (changing at each iteration). *Oblivious* dimension reduction techniques relieve us from the need to adapt the procedure to each of these subspaces. Our *modewise* embedding operator $L$ acts without initial vectorization of a tensor and results in $n^d/(m^d + dmn)$ $(m \ll n)$ memory reduction:

$$L(\mathcal{X}) := \mathbf{A_0}\left(\text{vectorize}\left(\mathcal{X} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\right)\right), \quad \text{where } \times_j \text{ is a } j\text{-mode product.} \tag{1}$$

Fig. 1 illustrates our compression process. We give the analysis for the cases (a) when matrices $\mathbf{A_k}$ are taken from a general class of $\eta$-optimal JL embedding distributions, which includes random matrices with i.i.d. subgaussian entries, as well as sparse JL constructions and others, and (b) in the case when $\mathbf{A_k}$ have special Kronecker Fourier JL form (introduced in the recent paper [JKW19]). In the general case (a) we prove

**Theorem 3.** *Let $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ and $\mathcal{L}$ be an $r$-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ spanned by rank-one tensors, which component vectors are sufficiently incoherent.*

*Then, for any $m' \le Cr \cdot \varepsilon^{-2} \cdot \ln(47/\varepsilon\sqrt[d]{\eta})$ a linear operator $L : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m'}$ as per (1) satisfies the following with probability at least $1 - \eta$*

$$\left| \|L(\mathcal{X} - \mathcal{Y})\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right| \le \varepsilon \|\mathcal{X} - \mathcal{Y}\|^2 \quad \text{for all } \mathcal{Y} \in \mathcal{L}. \tag{2}$$

In the latter case (b) we achieve the same target dimension up to log-factors, but do not need incoherence and rank-one basis tensor assumptions. Also, the intermediate dimension (that we achieve after the modewise products before the vectorization) becomes especially efficient, $O_{\log,d}(r^2\varepsilon^{-2})$.

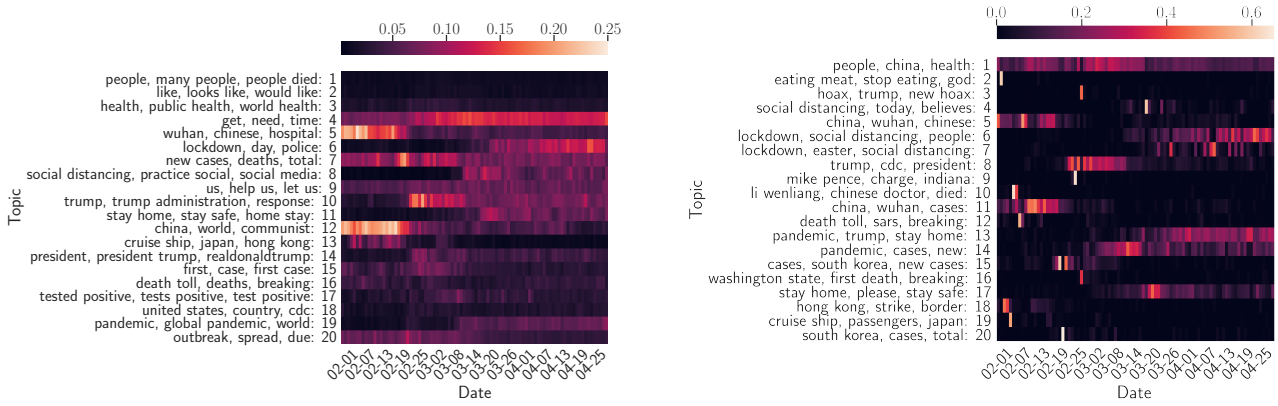


FIGURE 3. Tensor methods (NCPD, right) pick up short-term topics and attribute them to the correct dates, matrix methods detect global trends only (NMF, left); COVID-19 Twitter data: 1000 most retweeted tweets/day (Feb-May 2020)

**Outlook:** The data in the real applications is often multi-modal: for example, a 2-d picture usually has the third dimension representing color, and a movie has yet another temporal component. The tensor structure is inherent for such data, however, the majority of existing tensor methods (e.g. [LHW17, WTSA15, SGTU18]) essentially disregard it, applying matrix (or vector) techniques to matricizations (or vectorizations) of the tensors. In addition to the loss of structure, this approach results in huge memory and computational requirements even for tensors with compact structures. The development of the *modewise methods for tensors* is one of the key directions of my research. My ongoing projects include (a) proving a tensor version of the restricted isometry property [CT05] to guarantee a possibility of tensor recovery from a few modewise measurements and (b) developing modewise versions of the polynomial kernel sketching methods ([AKK$^+$20, ANW14, PP13]).

I also work on creating *interpretable machine learning techniques* using tensor methods. Low-rank tensor (and matrix) decompositions identify intrinsic components ("topics") in the data. Imposing non-negativity on the factors makes the topics interpretable. In [KKL$^+$20], my colleagues and I used non-negative CP tensor decomposition (NCPD [CC70, H$^+$70]) for dynamic topic modeling on Twitter text data related to the COVID-19 pandemic. We were able to discover a variety of related topics (including political events, personal beliefs about COVID-19 and calls to action), both persistent and short-term, and successfully attribute them to the days they were trending. Relative to its matrix counterparts, NCPD captures the topic structure more precisely and are considerably better at detecting smaller topics, see also Fig. 3. The next goal is to speed up NCP fitting using modewise dimension reduction techniques. Multiple other directions related to the topic-aware learning from data include (a) adding flexibility to these originally low parametric methods (unlike famous neural networks, vanilla NCPD has just one parameter – number of topics – and benefits a lot from a proper regularization) (b) extension of the related supervised methods [LYC09] to incorporate side information about the topics or feature importance, (c) topic-aware data search.

## 3. Linear systems: Randomized iterative linear solvers for error correction

One of the most ubiquitous problems arising across the sciences is that of solving large-scale systems of linear equations, $Ax = b$. Scalable and efficient iterative methods are used when it is too slow or infeasible to solve the system directly by inversion. Iterative methods frequently employ randomization: for example, in the famous stochastic gradient descent method (SGD, [Bot10]) gradients are approximated based on random subsamples of the data (which speeds up the iterations significantly). In the randomized Kaczmarz method[1](RK, [SV09]), random choice of the next step helps to avoid a malicious or unlucky ordering of equations that might lead to slow convergence. Moreover, it allows one to use high-dimensional

[1]RK iteratively projects each current approximation $x_k$ onto the solution space of the next *randomly chosen* equation of the full column rank overdetermined linear system until convergence.

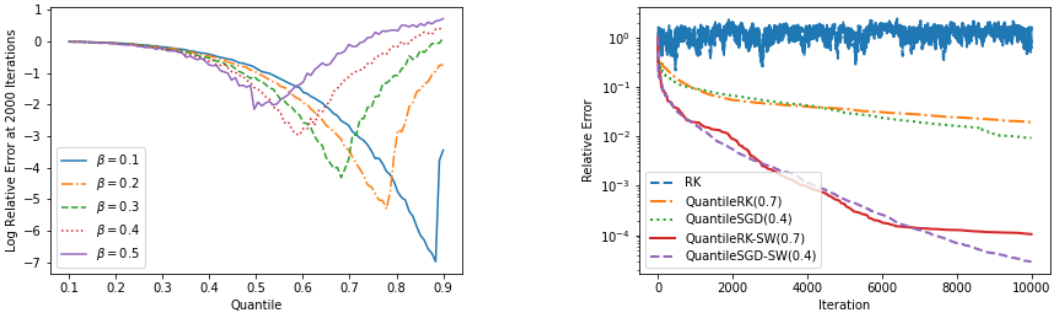


FIGURE 4. QuantileRK(q) works on up to 50% corruption rate, given proper $q$, $50000 \times 100$ Gaussian system (left); Performance of quantile-based methods on $699 \times 10$ Wisconsin Breast Cancer dataset, $\beta = 0.143$ (right)

probability methods to prove that RK converges exponentially in expectation, namely,

$$\mathbb{E}\|x_k - x_*\|_2^2 \le \left(1 - R^{-2}\right)^k \|x_0 - x_*\|_2^2, \quad \text{where} \quad R = \|A\|_F / \sigma_{min}(A). \tag{3}$$

where $x_*$ is the solution of the overdetermined system $Ax = b$. There are plenty of extensions of RK, speeding it up, aiming to alleviate the effect of row coherence that makes iterations of RK much less efficient, and extending it to the *inconsistent systems* (e.g., [NW13, Nee10, NT14], including my recent analysis of block Gaussian Kaczmarz method [RN20], joint with D. Needell). A standard way to treat the inconsistent (or, *noisy*) case $Ax^* = b + e$ is to show that the iterates approach the least squares solution $\hat{x} = arg\min_x \|Ax - b\|_2^2$ and quantify the distance to the true solution $x^*$ in terms of the noise size $\|e\|$. However, this is not satisfactory for the case of large and potentially adversarial *corruptions* in the vector $b$: the methods themselves should be modified to *avoid corrupted equations*.

In my work with J.Haddock, D.Needell and W.Swartsworth [HNRS20], we address this problem by proposing versions of RK and SGD methods that use order statistics of the residual (that is, distances from $x_k$ to the solution hyperplanes of the individual equations) to judge whether the next attempted iteration is safe. QuantileRK is a "lazy" algorithm which makes the attempted step only if it is safe, and QuantileSGD defines a safe step size based on the quantiles of the residual (see [HNRS20] for the algorithm details). We prove the following.

**Theorem 4.** *Let the matrix $A$ have subgaussian isotropic rows, with the entries that have centered and bounded density functions. Then with probability $1 - ce^{-c_q m}$, the iterates produced by the QuantileRK(q) and QuantileSGD(q) converge with the standard convergence rate (3) with $R = C\|A\|_F/\sigma_{min}(A)$ (same order as the Kaczmarz rate for uncorrupted systems) as long as the fraction of corrupted entries $\beta$ is small (we put no restrictions on their magnitude), $q \le 1/2 - \beta$, and $m \gg n \log n$.*

Theoretical analysis in [RN20, HNRS20] is based on both known and novel probabilistic concentration of measure results. Experimentally, we see that our methods work on up to 50% of incoherent corruptions, and up to 20% of adversarial corruptions (that consistently create an "alternative" solution of the system). Preliminary results on real-world datasets are also available, see Fig. 4 and [HNRS20].

**Outlook:** Numerous applications face a need to solve large-scale systems involving corrupted measurements, ranging from medical imaging and sensor networks to error correction and data science. In my future work, I plan to develop methods for the more general corruption models, in which there is a mix of non-sparse noise and sparse, but large, corruptions on the vector $b$ or the case when the noise (or corruptions) additionally affects the matrix $A$. One application of the latter is the low rank tensor fitting algorithm (discussed in Section 2) with corrupted measurements.

Finally, solving linear systems is only a starting point: I plan to develop and analyze randomized iterative methods that avoid significantly large unknown corruptions for a general class of optimization problems (such as the systems with non-linearities and inequalities, finding extrema of functions and beyond), which has applications in training machine learning models in the presence of adversarial training data.

## References


[AKK+20] Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 141–160. SIAM, 2020.

[ANW14] Haim Avron, Huy Nguyen, and David Woodruff. Subspace embeddings for the polynomial kernel. In *Advances in neural information processing systems*, pages 2258–2266, 2014.

[BKY17] Roland Bauerschmidt, Antti Knowles, and Horng-Tzer Yau. Local semicircle law for random regular graphs. *Communications on Pure and Applied Mathematics*, 70(10):1898–1960, 2017.

[BL13] Shimon Brooks and Elon Lindenstrauss. Non-localization of eigenfunctions on large regular graphs. *Israel Journal of Mathematics*, 193(1):1–14, 2013.

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[CC70] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[CT05] Emmanuel J Candés and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[Dem97] James W Demmel. *Applied numerical linear algebra*. SIAM, 1997.

[Ede88] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. A.*, 9(4):543–560, 1988.

[H+70] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. 1970.

[Hås90] Johan Håstad. Tensor rank is NP-complete. *Journal of algorithms (Print)*, 11(4):644–654, 1990.

[HNRS20] J. Haddock, D. Needell, E. Rebrova, and W. Swartworth. Quantile-based iterative methods for corrupted systems of linear equations. 2020. Submitted.

[INRZ19] MA Iwen, Deanna Needell, Elizaveta Rebrova, and Ali Zare. Lower memory oblivious (tensor) subspace embeddings with fewer random bits: Modewise methods for least squares. *arXiv preprint arXiv:1912.08294*, 2019.

[JKW19] Ruhui Jin, Tamara G Kolda, and Rachel Ward. Faster Johnson-Lindenstrauss transforms via Kronecker products. *arXiv preprint arXiv:1909.04801*, 2019.

[KB09] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[KKL+20] L. Kassab, A. Kryshchenko, H. Lyu, D. Molitor, D. Needell, and E. Rebrova. On nonnegative matrix and tensor decompositions for COVID-19 Twitter dynamics. 2020. Submitted.

[LHW17] Xingguo Li, Jarvis Haupt, and David Woodruff. Near optimal sketching of low-rank tensor regression. In *Advances in Neural Information Processing Systems*, pages 3466–3476, 2017.

[LYC09] Hyekyoung Lee, Jiho Yoo, and Seungjin Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2009.

[Nee10] Deanna Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT*, 50(2):395–403, 2010.

[NT14] Deanna Needell and Joel A Tropp. Paved with good intentions: analysis of a randomized block Kaczmarz method. *Linear Algebra Appl.*, 441:199–221, 2014.

[NW13] Deanna Needell and Rachel Ward. Two-subspace projection method for coherent overdetermined systems. *J. Fourier Anal. Appl.*, 19(2):256–269, 2013.

[PP13] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247, 2013.

[Reb19] Elizaveta Rebrova. Constructive regularization of the random matrix norm. *Journal of Theoretical Probability*, pages 1–23, 2019.

[RN20] Elizaveta Rebrova and Deanna Needell. On block Gaussian sketching for the Kaczmarz method. *Numerical Algorithms*, pages 1–31, 2020.

[RT18] Elizaveta Rebrova and Konstantin Tikhomirov. Coverings of random ellipsoids, and invertibility of matrices with iid heavy-tailed entries. *Israel J. Math.*, 227(2):507–544, 2018.

[RV08] Mark Rudelson and Roman Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008.

[RV16] Mark Rudelson and Roman Vershynin. No-gaps delocalization for general random matrices. *Geometric and Functional Analysis*, 26(6):1716–1776, 2016.

[RV18] Elizaveta Rebrova and Roman Vershynin. Norms of random matrices: local and global problems. *Adv. Math.*, 324:40–83, 2018.

[SGTU18] Yiming Sun, Yang Guo, Joel A Tropp, and Madeleine Udell. Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*, 2018.

[Sil89] Jack W Silverstein. On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 30(2):307–311, 1989.

[SV09] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[TV08] Terence Tao and Van Vu. Random matrices: the circular law. *Communications in Contemporary Mathematics*, 10(02):261–307, 2008.

[TV10] Terence Tao and Van Vu. Smooth analysis of the condition number and the least singular value. *Mathematics of Computation*, 79(272):2333–2352, 2010.

[TV17] Oguzhan Teke and Palghat P Vaidyanathan. Uncertainty principles and sparse eigenvectors of graphs. *IEEE Transactions on Signal Processing*, 65(20):5406–5420, 2017.

[Ver16] Roman Vershynin. High-dimensional probability. *An Introduction with Applications*, 2016.

[WTSA15] Yining Wang, Hsiao-Yu Tung, Alexander J Smola, and Anima Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *Advances in Neural Information Processing Systems*, pages 991–999, 2015.