Sketching for Motzkin's Iterative Method for Linear Systems

Elizaveta Rebrova[*,1] and Deanna Needell[*,2]

[*]Department of Mathematics, University of California, Los Angeles, Los Angeles, California
[1]rebrova@math.ucla.edu
[2]deanna@math.ucla.edu

Abstract— **Projection-based iterative methods for solving large over-determined linear systems are well-known for their simplicity and computational efficiency. It is also known that the correct choice of a sketching procedure (i.e., preprocessing steps that reduce the dimension of each iteration) can improve the performance of iterative methods in multiple ways, such as, to speed up the convergence of the method by fighting inner correlations of the system, or to reduce the variance incurred by the presence of noise. In the current work, we show that sketching can also help us to get better theoretical guarantees for the projection-based methods. Specifically, we use good properties of Gaussian sketching to prove an accelerated convergence rate of the sketched relaxation (also known as Motzkins) method. The new estimates hold for linear systems of arbitrary structure. We also provide numerical experiments in support of our theoretical analysis of the sketched relaxation method.**

I. INTRODUCTION

We are interested in solving large-scale systems of linear equations

$$Ax = b, \qquad (I.1)$$

where $b \in \mathbb{R}^m$ and A is an $m \times n$ matrix. We will consider the case when $m \geq n$, when iterative methods are typically employed.

One of the most popular iterative methods is the Kaczmarz method, being simple, efficient and well-adapted to large amounts of data. The original Kaczmarz method starts with some initial guess x_0, and then iteratively projects the previous approximation x_k onto the solution space of the next equation in the system. Namely, if $a_1, \ldots, a_m \in \mathbb{R}^n$ are the rows of A, then the k-th step of the algorithm is

$$x_k = x_{k-1} + \frac{b_i - \langle a_i, x_{k-1} \rangle}{\|a_i\|^2} a_i, \qquad (I.2)$$

where $b = (b_1, \ldots, b_n) \in \mathbb{R}^m$ is the right hand side of the system, $x_{k-1} \in \mathbb{R}^n$ is the approximation of a solution x_* obtained in the previous step. To provide theoretical guarantees for the convergence of the method, Strohmer and Vershynin [14] proposed to choose the next row

index i at random with the probabilities weighted proportionally to the L_2 norms of the rows a_i. The authors have shown that this *randomized Kaczmarz algorithm* is guaranteed to converge exponentially in expectation, namely,

$$\mathbb{E}\, \|x_k - x_*\|_2^2 \leq \left(1 - \frac{1}{\tilde{\kappa}(A)}\right)^k \|x_0 - x_*\|_2^2, \quad (I.3)$$

where $\tilde{\kappa}(A) = \|A\|_F^2 / s_{min}^2(A)$ is a condition number of the system and x_* is the solution of the system (I.1).

Here and further, $s_{min}(A)$ denotes the smallest singular value of the matrix A and $\|A\|_F := trace(\sqrt{A^*A})$ (Frobenius, or Hilbert-Shmidt, norm of the matrix). We always assume that the matrix A has full column rank, so that $s_{min}(A) > 0$.

However, the randomized scheme does not always choose the best equation on which to project. The approach by Agmon [1] and Motzkin and Schoenberg [9], resolves this issue. The idea is to always project onto the farthest equation from the current iteration point x_k, making the distance to the solution $\|x_{k+1} - x_*\|_2$ as good as possible by the next step. Precisely, the k-th step of the algorithm is defined by (I.2), where

$$i := \mathrm{argmax}_{[j]}(A_j x_k - b_j)^2. \qquad (I.4)$$

This so-called Motzkin's method has also been known as the Kaczmarz method with the most violated constraint or maximal-residual control ([13], [12], [2]).

It is intuitive to expect that Motzkin's method converges more efficiently than Kaczmarz *per iteration* (although each iteration is clearly much more expensive, as one has to swipe over all m equations to determine the best one to satisfy (I.4)) A step towards the theoretical guarantee of the convergence of Motzkin's method improved convergence rate was made in [7], where the authors have shown that

$$\mathbb{E}\, \|x_k - x_*\|_2^2 \leq \prod_{i=1}^{k-1}\left[1 - \frac{s_{min}^2(A)}{\gamma_i(A)}\right]^k \|x_0 - x_*\|_2^2, \ (I.5)$$

where $\gamma_i(A) := \|Ax_i - Ax_*\|_2^2/\|Ax_i - Ax_*\|_\infty^2$ is a so-called *dynamic range* of the system.

In general, to get a good estimate on the dynamic range, one would need to know more about the structure of the matrix A. The authors also give a heuristic for $\gamma_i(S)$ in the case when the matrix A has independent standard normal entries $\gamma_k \sim \|A\|_F^2 n/\log(m-k)$ ([7, Section 2.1]). So, it shows accelerated convergence for several initial iterations while $\log(m-k) > n$.

An independent branch of research improving Kaczmarz method was presented by Gower and Richtarik in [6]. The main idea of their *sketch-and-project* framework is the following. One can observe that a random selection of a row can be represented as a *sketch*, that is, left multiplication by a random vector (with exactly one non-zero entry). Then, the iteration (I.2) is a projection onto the image of the sketch. A natural way to extend the method is to pre-process every iteration by left multiplication with an $s \times m$ matrix S (or $1 \times m$ vector) taken from some distribution, thereby reducing the dimension of the problem (see, e.g., [6], [8], [10]).

II. PROPOSED METHOD

One could also use sketching to pre-process each iteration of Motzkin's method. Like in the original sketch-and-project framework, the idea is to reduce the dimension of the problem (in this case the space of search for the best equation in the sense of (I.4)), hopefully, preserving the main properties of the system.

Again, the k-th step of the algorithm is defined by (I.2), but we would choose the best row of the sketched matrix, namely.

$$i = argmax_{[i]}(S^T A)_i x_k - (S^T b)_i)^2 \qquad \text{(II.1)}$$

A. Block Sketched Motzkin (SKM)

A natural choice of the sketch is a *block sketch* (analogous to the one considered in [11] for the Kaczmarz algorithm): $m \times s$ sketches

$$S = (\text{zeros}(s, \text{shift}), I(s,s), \text{zeros}(s, m - \text{shift} - s))^T,$$

where zeros() denotes a matrix of all zeros, $I()$ the identity, and s is the block size and shift $= sz$, $z \in \{1, 2, ..., \lfloor m/s \rfloor\}$ is selected randomly at each step.

Then, application of the sketch is exactly equivalent to selecting an $s \times n$ row submatrix of A and proceeding with the Motzkin's scheme inside the selected submatrix. This method is equivalent to the SKM method studied in the more general case of inequalities in [4]. To the best of our knowledge, there are still no theoretical guarantees for the convergence of SKM (at least, unspecified to some model of the matrix A, such as a Gaussian model).

B. Gaussian Sketched Motzkin (GSM)

We propose to draw sketch matrices S from the Gaussian distribution on each step ($m \times s$ matrices with i.i.d. $N(0,1)$ entries independent between each other). In this framework, we can provide theoretical guarantees for the exponential convergence rate, *log*-accelerated with respect to the standard rate (I.3), without any assumptions about the model of the matrix A.

Theorem 2.1: Suppose A is a $m \times n$ matrix with full column rank ($m \geq n$) and let x_* be a solution of the system $Ax = b$. For any initial estimate x_0, the GSM method produces a sequence $\{x_k, k \geq 0\}$ of iterates that satisfy

$$\mathbb{E}\|x_k - x_*\|_2^2 \leq \left(1 - c\frac{\log s}{\tilde{\kappa}(A)}\right)^k \|x_0 - x_*\|_2^2. \quad \text{(II.2)}$$

Here, $c > 0$ is an absolute constant.

C. Sparse Gaussian sketched Motzkin (sGSM)

Comparing to SKM, the main disadvantage of GSM is the cost of each sketch: instead of drawing a sub-block of the original matrix A, one needs to pre-multiply A by a dense Gaussian matrix. As a trade-off between these two methods, we also propose the *sparse GSM* (or, sGSM) method, taking sketches

$$S := (\text{zeros}(s, \text{shift}), X, \text{zeros}(s, m-\text{shift}-s))^T, \quad \text{(II.3)}$$

where X is an $s \times s$ matrix with i.i.d. $N(0,1)$ entries. At each step, we generate a sketch matrix S independently from the previous steps. This method is equivalent to selecting a $s \times n$ row sub-block of A and then sketching only that sub-block by a Gaussian matrix X. Then, one can either choose the position of a sub-block to be sketched randomly on each step, or fix a specific well-conditioned sub-block of A and just resample the matrix X from the Gaussian distribution.

We can also provide theoretical guarantees for the convergence of the sGSM method.

Theorem 2.2: Suppose A is a $m \times n$ matrix with full column rank ($m \geq n$) and let x_* be a solution of the system $Ax = b$. For any initial estimate x_0, the sGSM method produces a sequence $\{x_k, k \geq 0\}$ of iterates that satisfy

$$\mathbb{E}\|x_k - x_*\|_2^2 \leq \left(1 - c\frac{\log s}{\tilde{\kappa}(A')}\right)^k \|x_0 - x_*\|_2^2. \quad \text{(II.4)}$$

Here, $c > 0$ is an absolute constant and A' has maximal $\tilde{\kappa}(A')$ over all $s \times n$ row submatrices of A.

Remark 2.3: If we know the condition number $\tilde{\kappa}'$ of a $s \times n$ row block that we use in iterations, Theorem II.4 holds with $\tilde{\kappa}'$ in place of the worst case $\tilde{\kappa}(A')$. There

is an abundant literature on existence and construction of well-conditioned sub-blocks for any matrix (see [16], [18], [17] as well as the discussion in [11]). Also, in many standard cases a random selection of a sub-block produces a well-conditioned submatrix (see [15], [3]). For our experiments in Section III we split the matrix into $\lfloor m/s \rfloor$ blocks of the same size and every time pick a sub-block randomly with replacement.

D. Proof of the theoretical results

Proposition 2.4: Suppose A is a $m \times n$ matrix with full column rank ($m \geq n$) and let x_* be a solution of the system $Ax = b$. Let x_k be a fixed vector in $\mathbb{R}^n$. Let S be $m \times s$ matrix with i.i.d. standard normal entries and x_{k+1} is obtained by iteration (I.2) with the index i chosen by the rule (II.1). Then,

$$\mathbb{E}_S \|x_{k+1} - x_*\|_2^2 \leq (1 - c\frac{\log s \cdot \sigma_{min}^2(A)}{\|A\|_F^2})\|x_k - x_*\|_2^2.$$

Proof: Using that $x_{k+1} - x_*$ is orthogonal to $(S^T A)_i$, where i is the index chosen for the k-th iteration, we have the following by Pythagorean theorem:

$$\|x_{k+1} - x_*\|_2^2 = \|x_k - x\|_2^2 - \frac{\|(S^T b)_i - (S^T A)_i x_k\|_2^2}{\|(S^T A)_i\|_2^2}$$
$$= \|x_k - x\|_2^2 - \frac{\|S^T b - S^T A x_k\|_\infty^2}{\|(S^T A)_i\|_2^2}.$$

since $i = \operatorname{argmax}((S^T b)_i - (S^T A)_i x_k)^2$.

Now, since we consider the previous iteration fixed (so, x_k and i are fixed),

$$\mathbb{E}_S \|x_{k+1} - x_*\|_2^2 = \|x_k - x\|_2^2 - \mathbb{E}\frac{\|S^T b - S^T A x_k\|_\infty^2}{\|(S^T A)_i\|_2^2}.$$

Using that the function $(x, y) \mapsto x^2/y$ is convex on the positive orthant, by Jensen's inequality, we can estimate

$$\mathbb{E}\frac{\|S^T b - S^T A x_k\|_\infty^2}{\|(S^T A)_i\|_2^2} \geq \frac{(\mathbb{E}_S \|S^T b - S^T A x_k\|_\infty)^2}{\mathbb{E}(\|(S^T A)_i\|_2^2)}$$
$$= \frac{(\mathbb{E}_S \|S^T A(x_* - x_k)\|_\infty)^2}{\|A\|_F^2},$$

where in the denominator we computed

$$\mathbb{E}\|(S^T A)_i\|_2^2 = \sum_{k=1}^n \mathbb{E}(\sum_{j=1}^m S_{ij} A_{jk})^2 = \sum_{kj} A_{jk}^2 = \|A\|_F^2.$$

Moreover, from the estimate for the maximum of independent normal random variables,

$$\mathbb{E}_S \|S^T A(x_* - x_k)\|_\infty = \mathbb{E}\max_{i \in [s]}\langle S_i, v \rangle \geq c\|v\|_2\sqrt{\log s},$$

where $v = A(x - x_k)$, then

$$\mathbb{E}_S \|x_{k+1} - x_*\|_2^2 \leq \|x_k - x\|_2^2 - \frac{c\|v\|_2^2 \log s}{\|A\|_F^2}$$
$$\leq \|x_k - x\|_2^2 - \frac{c\sigma_{min}^2(A) \log s}{\|A\|_F^2}\|x_k - x\|_2^2.$$

This concludes the proof of the Proposition 2.4. ∎

Proof of Theorem 2.1. We have

$$\mathbb{E}\|x_k - x_*\|_2^2 = \mathbb{E}_{S_1}\mathbb{E}_{S_2}\ldots\mathbb{E}_{S_k}\|x_k - x_*\|_2^2$$
$$\leq \left[1 - c\frac{\log s \cdot \sigma_{min}^2(A)}{\|A\|_F^2}\right]^k \|x_0 - x_*\|_2^2.$$

Here, $\mathbb{E}_{S_1}, \ldots, \mathbb{E}_{S_k}$ refer to the randomness of choosing a matrix S_i (independent from all Gaussian sketches $S_1, \ldots, S_{i-1}$ that were used during previous steps). The last inequality is guaranteed by Proposition 2.4. This concludes the proof of Theorem 2.1.

Proof of Theorem 2.2. Note that for the sparse Gaussian sketch defined by (II.3), $S^T A = X^T A'$, where A' is an $s \times n$ row sub-block of A. Proposition 2.4 holds for $A = A'$ and $S = X$. Then, the proof is concluded along the lines of the proof of Theorem 2.1.

III. EXPERIMENTAL DATA AND RESULTS

In this section, we present some numerical experiments on the performance of the Sketched Motzkin methods. Everything was coded and run in MATLAB R2019a, on a 1.6GHz dual-core Intel Core i5, 8 GB 2133 MHz. Time is always measured in seconds. We generate the solution of a system as a random vector (and define the left hand side b accordingly), so we do not need to worry about the case when $\|x_*\|_2 = 0$. We use $x_0 = 0$ as an initial point.

The first two figures show per iteration performance on two main models of the matrix A: an incoherent Gaussian matrix with i.i.d. $N(0, 1)$ entries and a coherent matrix with almost collinear rows, $A_{ij} \sim \text{Unif}[0.8, 1]$.

As expected, Gaussian sketching is as good as block sketching (SKM) when the rows of A are Gaussian themselves, so SKM, GSM and sGSM show the same performance on the Gaussian model (Figure 1). Also not surprisingly, choosing the most violated equation (Motzkin) shows the best per iteration convergence progress, whereas taking the next equation at random (Kaczmarz) is the weakest.

We can see in Figure 2 that Gaussian sketching significantly improves the performance of Motzkin's method on the coherent model, and that lighter sparse sketches (sGSM) work as well as the dense GSM in practice.

In Figure 3, we can see that this per iteration advantage is still not enough for Gaussian sketched methods to show the best performance in time. Still, it is
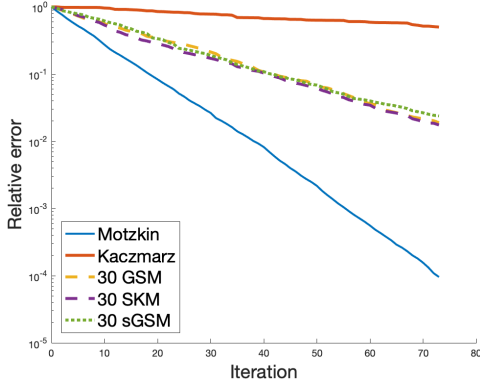
3

Fig. 1: Performance of the methods on $A = 5000 \times 100$ matrix with i.i.d. $N(0,1)$ entries
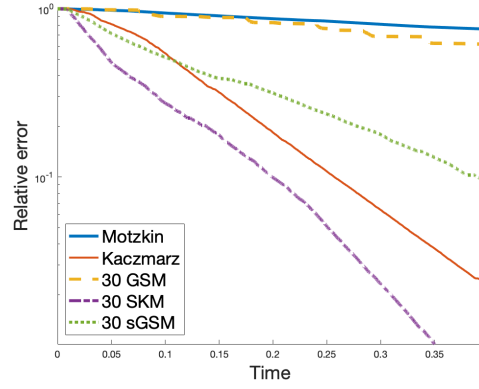

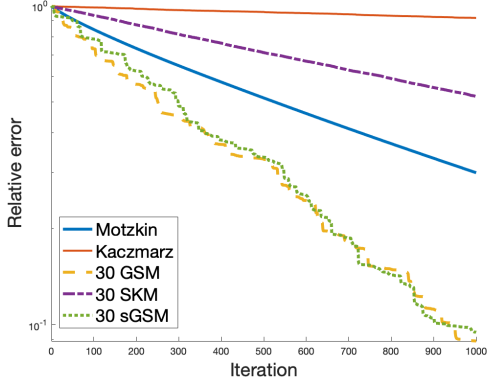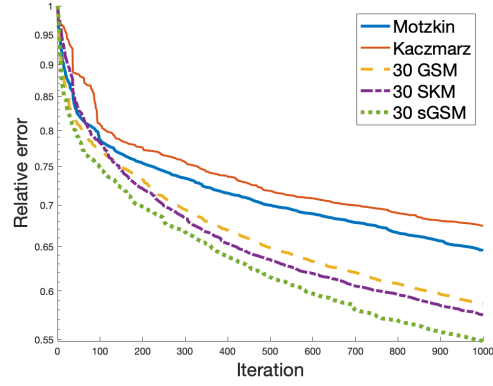Fig. 2: Performance of the methods on $A = 5000 \times 100$ matrix with i.i.d. $Unif[0.8., 1]$ entries

practical to use sketching: the SKM method gives the best convergence in time. However, with the use of distributed computations or fast matrix multiplication one potentially could speed up the application of the Gaussian sketches, making GSM and sGSM methods more efficient. We can see in Figure 3 the advantage of using sparse Gaussian sketches rather than dense GSM.

We also compare the preformance of the methods discussed above on two real world datasets, GAS and COVTYPE, both taken from the UCI repository [5].

Finally, we compare various sketch sizes for Motzkin's method. As we can see in Figure 6, increasing the block size improves per iteration performance (which is expected since we keep more and more information about the system). Clearly, bigger blocks mean slower iterations, so the question about an optimal block size requires the comparison in time. In Figure 7, we compare the time required to achieve certain error threshold for
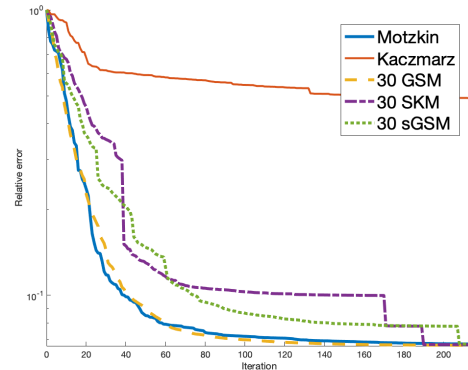

Fig. 3: Performance in time on $A = 5000 \times 100$ matrix with i.i.d. $Unif[0.8., 1]$ entries

Fig. 4: Performance of the methods on a real-world dataset, GAS (1000×128)

Fig. 5: Performance of the methods on a real-world dataset, COVTYPE (5000×54)
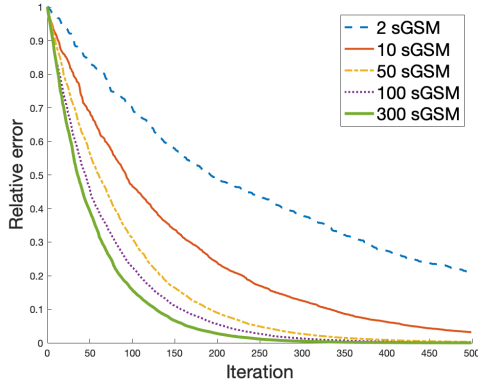
Fig. 6: Dependence of the convergence rate on the block size. $A = 5000 \times 100$ matrix with i.i.d. $N(0, 1)$ entries
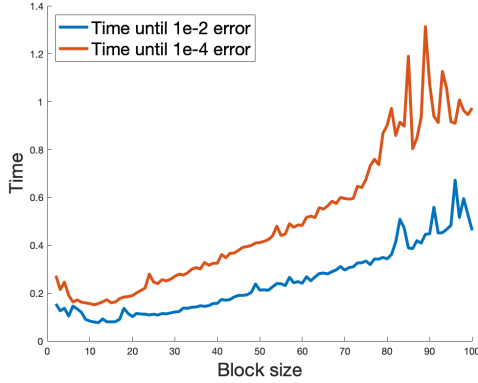

Fig. 7: sGSM method: $A = 5000 \times 100$ i.i.d. matrix from $Unif[0.8, 1]$ model; time is averaged over 20 iterations

sGSM methods, varying the block size s from one to $n = 100$. We observe that there is a non-trivial optimal block size (somewhere between 5 and 20 in our experiments).

IV. Conclusions

In this paper, we analyzed the application of sketching to Motzkin's iterative method for solving consistent overdetermined large system of equations. We considered three ways to sketch Motzkin's method: SKM, GSM and sparseGSM. We provide theoretical guarantees for the accelerated convergence of GSM (and sparseGSM for a well-conditioned matrix). In the experiments section, we have shown some cases when sketched methods work better than both Kaczmarz and Motzkin (and when Gaussian sketches outperform SKM method). Finally, we investigate experimentally optimal block size for the sparseGSM method. One of the interesting future directions of the current work would be to investigate the dependence of the optimal block size on the model and provide theoretical estimates for it.

References

[1] S. Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954.

[2] Y. Censor. Row-action methods for huge and sparse systems and their applications. *SIAM review*, 23(4):444–466, 1981.

[3] S. Chrétien and S. Darses. Invertibility of random submatrices via tail-decoupling and a matrix chernoff inequality. *Statistics & Probability Letters*, 82(7):1479–1487, 2012.

[4] J. A. De Loera, J. Haddock, and D. Needell. A sampling Kaczmarz–Motzkin algorithm for linear feasibility. *SIAM Journal on Scientific Computing*, 39(5):S66–S87, 2017.

[5] D. Dua and C. Graff. UCI machine learning repository, 2017.

[6] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[7] J. Haddock and D. Needell. On Motzkins method for inconsistent linear systems. *BIT Numerical Mathematics*, 59(2):387–401, 2019.

[8] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

[9] T. S. Motzkin and I. J. Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:393–404, 1954.

[10] D. Needell and E. Rebrova. On block gaussian sketching for iterative projections. *arXiv preprint arXiv:1905.08894*, 2019.

[11] D. Needell and J. A. Tropp. Paved with good intentions: analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.

[12] J. Nutini, B. Sepehry, I. Laradji, M. Schmidt, H. Koepke, and A. Virani. Convergence rates for greedy kaczmarz algorithms, and faster randomized kaczmarz rules using the orthogonality graph. *arXiv preprint arXiv:1612.07838*, 2016.

[13] S. Petra and C. Popa. Single projection kaczmarz extended algorithms. *Numerical Algorithms*, 73(3):791–806, 2016.

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

[15] J. A. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346(23-24):1271–1274, 2008.

[16] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

[17] L. Tzafriri and J. Bourgain. On a problem of kadison and singer. *Journal für die reine und angewandte Mathematik*, 420:1–44, 1991.

[18] R. Vershynin et al. Random sets of isomorphism of linear operators on hilbert space. In *High dimensional probability*, pages 148–154. Institute of Mathematical Statistics, 2006.