Norms of random submatrices and sparse approximation

Joel A. Tropp [1]

Applied & Computational Mathematics, California Institute of Technology, Pasadena, CA 91125-5000.

Received *****; accepted after revision +++++

Presented by ?????

Abstract

Many problems in the theory of sparse approximation require bounds on operator norms of a random submatrix drawn from a fixed matrix. The purpose of this note is to collect estimates for several different norms that are most important in the analysis of ℓ_1 minimization algorithms. Several of these bounds have not appeared in detail.

Résumé

Sur la norme de sous-matrice tirée aléatoirement. Beaucoup de problèmes en théorie d'approximation non linéaire demandent de majorer la norme d'une matrice aléatoirement extraite d'une matrice fixe de plus grandes dimensions. L'objectif de cette note est de présenter quelques estimations de ces normes qui se revèlent être importantes pour l'étude des algorithmes de minimisation de type ℓ_1. Plusieurs de ces bornes n'ont pas encore été publiées explicitement.

1. Introduction

We consider matrices written with respect to the standard basis, and we focus on three specific norms. The norm $\|\cdot\|$ is the usual Hilbert space operator norm; the ℓ_1 to ℓ_2 operator norm $\|\cdot\|_{1\to 2}$ computes the maximum ℓ_2 norm of a column; and $\|\cdot\|_{\max}$ returns the maximum absolute entry of a matrix. Throughout, $\{\delta_j\}$ is a sequence of independent 0–1 random variables with common mean δ. We write $\boldsymbol{R}$ for the square diagonal matrix whose jth diagonal entry is δ_j; the dimensions of $\boldsymbol{R}$ are determined by context. The symbol $\mathbb{E}_p$ indicates the L_p norm of a random variable, i.e., $\mathbb{E}_p X = (\mathbb{E}\,|X|^p)^{1/p}$.

The main theorem is a bound on the spectral norm of a random principal submatrix.

Email address: `jtropp@acm.caltech.edu` (Joel A. Tropp).

URL: `http ://www.acm.caltech.edu/ jtropp` (Joel A. Tropp).

[1] This work was supported in part by DARPA/ONR N660010612011.

Theorem 1.1 (Random principal submatrices) *Let $\boldsymbol{A}$ be an $n \times n$ Hermitian matrix, decomposed into diagonal and off-diagonal parts: $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{H}$. Fix p in $[2, \infty)$, and set $q = \max\{p, 2\log n\}$. Then*

$$\mathbb{E}_p \|\boldsymbol{RAR}\| \le \mathrm{C}\left[q\,\mathbb{E}_p \|\boldsymbol{RHR}\|_{\max} + \sqrt{\delta q}\,\mathbb{E}_p \|\boldsymbol{HR}\|_{1,2} + \delta \|\boldsymbol{H}\|\right] + \mathbb{E}_p \|\boldsymbol{RDR}\|.$$

A partial case of this theorem appears in [5]. The argument is based on [4] and classical ideas from [3]. We apply the result to sparse approximation in Section 5. From this moment bound, tail probabilities can be estimated by applying Markov's inequality in the usual fashion.

2. Preliminaries

We begin with some background. First, we present a decoupling result for the spectral norm that refines a classical proposition from harmonic analysis [1].

Proposition 2.1 (Decoupling) *Let $\boldsymbol{H}$ be an Hermitian matrix with a zero diagonal. Then*

$$\mathbb{E}_p \|\boldsymbol{RHR}\| \le 2\,\mathbb{E}_p \|\boldsymbol{RHR}'\|$$

where the two random restrictions on the right-hand side are independent and identically distributed.

Proof. We establish the result for $p = 1$. Let $\boldsymbol{H}_{jk}$ be the matrix with entry h_{jk} in position (j, k) and zero elsewhere. Let η_j be iid 0–1 random variables with mean 1/2. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{RHR}\| &= \mathbb{E}\left\|\sum\nolimits_{j<k} \delta_j\delta_k(\boldsymbol{H}_{jk} + \boldsymbol{H}_{kj})\right\| \\ &\le 2\,\mathbb{E}_{\boldsymbol{\eta}}\,\mathbb{E}_{\boldsymbol{\delta}}\left\|\sum\nolimits_{j<k} [\eta_j(1-\eta_k) + \eta_k(1-\eta_j)]\,\delta_j\delta_k(\boldsymbol{H}_{jk} + \boldsymbol{H}_{kj})\right\|.\end{aligned}$$

There is a 0–1 vector $\boldsymbol{\eta}^\star$ for which the expression exceeds its expectation over $\boldsymbol{\eta}$. Let $T = \{j : \eta_j^\star = 1\}$.

$$\mathbb{E}\|\boldsymbol{RHR}\| \le 2\,\mathbb{E}\left\|\sum\nolimits_{\substack{j\in T\\ k\in T^c}} \delta_j\delta_k(\boldsymbol{H}_{jk} + \boldsymbol{H}_{kj})\right\| = 2\,\mathbb{E}\left\|\sum\nolimits_{\substack{j\in T\\ k\in T^c}} \delta_j\delta_k\boldsymbol{H}_{jk}\right\| = 2\,\mathbb{E}\left\|\sum\nolimits_{\substack{j\in T\\ k\in T^c}} \delta_j\delta_k'\boldsymbol{H}_{jk}\right\|.$$

where $\{\delta_k'\}$ is an independent copy of the sequence $\{\delta_j\}$. The first equality follows from a standard identity for block counter-diagonal Hermitian matrices. Now, the norm of a submatrix does not exceed the norm of the matrix, so we re-introduce the missing entries to complete the argument.

$$\mathbb{E}\|\boldsymbol{RHR}\| \le 2\,\mathbb{E}\left\|\sum\nolimits_{j\ne k} \delta_j\delta_k'\boldsymbol{H}_{jk}\right\| = 2\,\mathbb{E}\|\boldsymbol{RHR}'\|. \qquad \square$$

We also need a novel re-coupling result. It is based on the same ideas, so we omit the proof.

Proposition 2.2 (Re-coupling) *Let $\boldsymbol{H}$ be an Hermitian matrix with a zero diagonal. Then*

$$\mathbb{E}_p \|\boldsymbol{RHR}'\|_{\max} \le 4\,\mathbb{E}_p \|\boldsymbol{RHR}\|_{\max}.$$

Third, we bound the expected maximum of a random subset of nonnegative scalars. See [4, Lemma 5.1] for related ideas.

Proposition 2.3 (Max of a random subset) *Let $a_1, a_2, \dots, a_n$ be nonnegative and $K = \lfloor \delta^{-1} \rfloor$. Then*

$$\mathbb{E}\max \delta_j a_j \le 2 \max_{|T|\le K} \frac{1}{K}\sum\nolimits_{j\in T} a_j \le \frac{2\delta}{1-\delta} \max_{|T|\le\delta^{-1}} \sum\nolimits_{j\in T} a_j.$$

Proof. We may take $\{a_j\}$ nonincreasing. The bound follows from a calculation and the fact $K \ge \delta^{-1} - 1$.

$$\mathbb{E}\max\delta_j a_j \le \mathbb{E}\sum\nolimits_{j=1}^K \delta_j a_j + a_{K+1} \le \delta\sum\nolimits_{j=1}^K a_j + \frac{1}{K}\sum\nolimits_{j=1}^K a_j \le \frac{2}{K}\sum\nolimits_{j=1}^K a_j. \qquad \square$$

3. Maximum column norm of a random submatrix

This section contains bounds on the maximum column norm of a matrix restricted to a random set of columns or a random set of rows. The first result is an easy application of Proposition 2.3.

Theorem 3.1 *Let $\boldsymbol{B}$ be an $m \times n$ matrix with columns $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$. When $p \geq 1$,*

$$\mathbb{E}_p \left\| \boldsymbol{B}\boldsymbol{R} \right\|_{1\to 2} \leq \frac{2\delta}{1-\delta} \max_{|T| \leq \delta^{-1}} \left[\sum\nolimits_{j \in T} \left\| \boldsymbol{b}_j \right\|_2^p \right]^{1/p}.$$

The second result is for random row restrictions. A partial case appears in [5, Prop. 13].

Theorem 3.2 *Let $\boldsymbol{B}$ be an $m \times n$ matrix. For p in $[2, \infty)$, set $q = \max\{p, 2\log n\}$. Then*

$$\mathbb{E}_p \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{1\to 2} \leq 2^{1.25}\sqrt{q}\, \mathbb{E}_p \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{\max} + \sqrt{\delta} \left\| \boldsymbol{B} \right\|_{1\to 2}.$$

The proof relies on a lemma that is established with Khintchine's inequality.

Lemma 3.3 *Let $\boldsymbol{X}$ be an $m \times n$ matrix. For $r \in [1, \infty)$, choose $q \geq \max\{r, 2\log n\}$. Then*

$$\mathbb{E}_r \max\nolimits_{k=1,2,\dots,n} \left| \sum\nolimits_{j=1}^m \varepsilon_j |x_{jk}|^2 \right| \leq 2^{0.25}\sqrt{q} \left\| \boldsymbol{X} \right\|_{\max} \left\| \boldsymbol{X} \right\|_{1\to 2}.$$

where $\{\varepsilon_j\}$ is a sequence of independent Rademacher variables.

Proof. First, we replace the maximum with the ℓ_q norm. Apply the inequalities of Jensen and Khintchine. Bound the sum over k by a maximum. Finally, apply Hölder's inequality:

$$\begin{aligned} \mathbb{E}_r \max\nolimits_k \left| \sum\nolimits_j \varepsilon_j |x_{jk}|^2 \right| &\leq \left[\mathbb{E} \left(\sum\nolimits_k \left| \sum\nolimits_j \varepsilon_j |x_{jk}|^2 \right|^q \right)^{r/q} \right]^{1/r} \leq \left[\sum\nolimits_k \mathbb{E} \left| \sum\nolimits_j \varepsilon_j |x_{jk}|^2 \right|^q \right]^{1/q} \\ &\leq \mathrm{C}_q \left[\sum\nolimits_k \left(\mathbb{E} \left| \sum\nolimits_j \varepsilon_j |x_{jk}|^2 \right|^2 \right)^{q/2} \right]^{1/q} \leq \mathrm{C}_q n^{1/q} \left[\max\nolimits_k \sum\nolimits_j |x_{jk}|^4 \right]^{1/2} \\ &\leq \mathrm{C}_q \mathrm{e}^{0.5} \max\nolimits_{j,k} |x_{jk}| \max\nolimits_k \left[\sum\nolimits_j |x_{jk}|^2 \right]^{1/2}. \end{aligned}$$

Finally, recall that the constant C_q from Khintchine's inequality is bounded by $2^{0.25}\mathrm{e}^{-0.5}\sqrt{q}$. □

Proof. (Theorem 3.2) Define $E = \mathbb{E}_p \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{1\to 2}$. Writing $r = p/2$, we elaborate the quantity E. Then we center the random variables and apply the usual symmetrization [3, Lem. 6.3]:

$$E^2 = \left[\mathbb{E} \left(\max\nolimits_k \sum\nolimits_j \delta_j |b_{jk}|^2 \right)^r \right]^{1/r} \leq 2 \left[\mathbb{E}_{\boldsymbol{\delta}} \mathbb{E}_{\boldsymbol{\varepsilon}} \left| \max\nolimits_k \sum\nolimits_j \varepsilon_j \delta_j |b_{jk}|^2 \right|^r \right]^{1/r} + \delta \left\| \boldsymbol{B} \right\|_{1\to 2}^2.$$

Invoke Lemma 3.3 with $\boldsymbol{X} = \boldsymbol{R}\boldsymbol{B}$. Afterward, Cauchy–Schwarz results in

$$E^2 \leq 2^{1.25}\sqrt{q} \left[\mathbb{E} \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{\max}^r \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{1\to 2}^r \right]^{1/r} + \delta \left\| \boldsymbol{B} \right\|_{1\to 2}^2 \leq 2^{1.25}\sqrt{q}\, \mathbb{E}_p \left\| \boldsymbol{R}\boldsymbol{B} \right\|_{\max} E + \delta \left\| \boldsymbol{B} \right\|_{1\to 2}^2.$$

Solutions to the relation $E^2 \leq \alpha E + \beta$ obey $E \leq \alpha + \sqrt{\beta}$. This point completes the proof. □

4. Spectral norms of random submatrices

The proof of Theorem 1.1 uses a result of Rudelson–Vershynin [4] to bound the spectral norm of a random column submatrix. Its proof is analogous with that of Theorem 3.2 but relies on a sharp noncommutative Khintchine inequality [2]. The explicit constant was obtained in [5, Prop. 12].

Theorem 4.1 (Rudelson–Vershynin) *Let $\boldsymbol{B}$ be an $m \times n$ matrix with rank r. For p in $[2, \infty)$, set $q = \max\{p, 2\log r\}$. Then*

$$\mathbb{E}_p \left\| \boldsymbol{B}\boldsymbol{R} \right\| \leq 3\sqrt{q}\, \mathbb{E}_p \left\| \boldsymbol{B}\boldsymbol{R} \right\|_{1\to 2} + \sqrt{\delta} \left\| \boldsymbol{B} \right\|.$$

Proof. (Theorem 1.1) Remove the matrix diagonal, then decouple the projectors with Proposition 2.1:

$$\mathbb{E}_p \|\boldsymbol{RAR}\| \le 2\,\mathbb{E}_p \|\boldsymbol{RHR}'\| + \mathbb{E}_p \|\boldsymbol{RDR}\|.$$

To estimate the first term, we apply the Rudelson–Vershynin theorem twice, once for for each projector:

$$\begin{aligned}\mathbb{E}_p \|\boldsymbol{RHR}'\| &\le 3\sqrt{q}\,\mathbb{E}_p \|\boldsymbol{RHR}'\|_{1\to 2} + \sqrt{\delta}\,\mathbb{E}_p \|\boldsymbol{R}'\boldsymbol{H}\| \\ &\le 3\sqrt{q}\,\mathbb{E}_p \|\boldsymbol{RHR}'\|_{1\to 2} + 3\sqrt{\delta q}\,\mathbb{E}_p \|\boldsymbol{HR}'\|_{1\to 2} + \delta\,\mathbb{E}_p \|\boldsymbol{H}\|.\end{aligned}$$

The maximum column norm bound, Theorem 3.2, yields

$$\mathbb{E}_p \|\boldsymbol{RHR}'\| \le 3\sqrt{q}\left[2^{1.25}\sqrt{q}\,\mathbb{E}_p \|\boldsymbol{RHR}'\|_{\max} + \sqrt{\delta}\,\mathbb{E}_p \|\boldsymbol{HR}'\|_{1\to 2}\right] + 3\sqrt{\delta q}\,\mathbb{E}_p \|\boldsymbol{HR}\|_{1\to 2} + \delta\,\mathbb{E}_p \|\boldsymbol{H}\|.$$

Since $\boldsymbol{R}'$ and $\boldsymbol{R}$ are identically distributed, we combine the second and third terms to reach

$$\mathbb{E}_p \|\boldsymbol{RAR}\| \le 15q\,\mathbb{E}_p \|\boldsymbol{RHR}'\|_{\max} + 12\sqrt{\delta q}\,\mathbb{E}_p \|\boldsymbol{HR}\|_{1\to 2} + 2\delta\,\mathbb{E}_p \|\boldsymbol{H}\| + \mathbb{E}_p \|\boldsymbol{RDR}\|.$$

Finally, apply the re-coupling result, Proposition 2.2, to the first term. □

5. Random subdictionaries

A *dictionary* is an $m \times n$ matrix $\boldsymbol{\Phi}$ whose columns have unit ℓ_2 norm. Define the hollow Gram matrix $\boldsymbol{H} = \boldsymbol{\Phi}^*\boldsymbol{\Phi} - \mathbf{I}$, and note that $\|\boldsymbol{H}\|_{1\to 2} < \|\boldsymbol{\Phi}^*\boldsymbol{\Phi}\|_{1\to 2} = \max_k \|\boldsymbol{\Phi}^*\boldsymbol{\varphi}_k\|_2 \le \|\boldsymbol{\Phi}\|$. A *random subdictionary* with expected cardinality δn is a column submatrix $\boldsymbol{\Phi}_T$ where $T = \{j : \delta_j = 1\}$.

The most important statistic associated with a dictionary is the *coherence* $\mu = \max_{j\neq k} |\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k\rangle|$. For a set T of columns, the *local 2-cumulative coherence* is the quantity

$$\mu_2(T) = \max\nolimits_{k\notin T}\left[\sum\nolimits_{j\in T} |\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k\rangle|^2\right]^{1/2}.$$

Theorem 3.2 allows us to estimate the local 2-cumulative coherence of a random subdictionary.

Corollary 5.1 *Let $T = \{j : \delta_j = 1\}$. When $p = 2\log n$, we have $\mathbb{E}_p\,\mu_2(T) \le 4\mu\sqrt{\log n} + \sqrt{\delta}\,\|\boldsymbol{\Phi}\|$.*

Proof. Observe that the local coherence $\mu_2(T) = \|\boldsymbol{RH}(\mathbf{I} - \boldsymbol{R})\|_{1\to 2} \le \|\boldsymbol{RH}\|_{1\to 2}$. Invoke Theorem 3.2 along with the facts $\|\boldsymbol{RH}\|_{\max} \le \mu$ and $\|\boldsymbol{H}\|_{1\to 2} < \|\boldsymbol{\Phi}\|$. □

We can use Theorem 1.1 to study the conditioning of a random subdictionary via the quantity $\|\boldsymbol{RHR}\|$.

Corollary 5.2 *For $p = 2\log n$, we have the bound*

$$\mathbb{E}_p \|\boldsymbol{RHR}\| \le \mathrm{C}\left[\mu \log n + \sqrt{\delta\,\|\boldsymbol{\Phi}\|^2 \log n}\right]. \tag{1}$$

Proof. Apply Theorem 1.1 with $\boldsymbol{A} = \boldsymbol{H}$, then introduce $\|\boldsymbol{RH}\|_{1\to 2} < \|\boldsymbol{\Phi}\|$ and $\|\boldsymbol{RHR}\|_{\max} \le \mu$. □

A subject for further investigation is to use Proposition 2.3 to sharpen the first term of the bracket in (1) when p is small. An elegant answer has remained elusive.

References

[1] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math*, 57(2):137–224, 1987.

[2] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Annalen*, 319:1–16, 2001.

[3] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, 1991.

[4] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, 54(4):Article 21, pp. 1–19, Jul. 2007.

[5] J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008.