A NOTE ON SUMS OF INDEPENDENT RANDOM MATRICES AFTER AHLSWEDE-WINTER

1. The method

Ashwelde and Winter [1] proposed a new approach to deviation inequalities for sums of independent random matrices. The purpose of this note is to indicate how this method implies Rudelson's sampling theorems for random vectors in isotropic position.

Let $X_1, \dots, X_n$ be independent random $d \times d$ real matrices, and let $S_n = X_1 + \cdots + X_n$. We will be interested in the magnitude of the deviation $\|S_n - \mathbb{E}S_n\|$ in the operator norm.

1.1. **Real valued random variables.**

Ashlwede-Winter's method [1] is parallel to the classical approach to deviation inequalities for real valued random variables. We briefly outline the real valued method. Let $X_1, \dots, X_n$ be independent mean zero random variables. We are interested in the magnitude of $S_n = \sum_i X_i$. For simplicity, we shall assume that $|X_i| \le 1$ a.s. This hypothesis can be relaxed to some control of the moments, precisely to having sub-exponential tail.

Fix a $t > 0$ and let $\lambda > 0$ be a parameter to be chosen later. We want to estimate

$$p := \mathbb{P}(S_n > t) = \mathbb{P}(e^{\lambda S_n} > e^{\lambda t}).$$

By Markov inequality and using independence, we have

$$p \le e^{-\lambda t}\mathbb{E}e^{\lambda S_n} = e^{-\lambda t} \prod_i \mathbb{E}e^{\lambda X_i}.$$

Next, Taylor's expansion and the mean zero and boundedness hypotheses can be used to show that, for every i,

$$\mathbb{E}e^{\lambda X_i} \lesssim e^{\lambda^2 \operatorname{Var} X_i}, \qquad 0 \le \lambda \le 1.$$

This yields

$$p \lesssim e^{-\lambda t + \lambda^2 \sigma^2}, \qquad \text{where} \quad \sigma^2 := \sum_i \operatorname{Var} X_i.$$

The optimal choice of the parameter $\lambda \sim \min(\tau/2\sigma^2, 1)$ implies Chernoff's inequality

$$p \lesssim \max\left(e^{-t^2/\sigma^2}, e^{-t/2}\right).$$

1.2. **Random matrices.** Now we try to generalize this method when $X_i \in \mathcal{M}_d$ are independent mean zero random matrices, where $\mathcal{M}_d$ denotes the class of symmetric $d \times d$ matrices.

Some of the matrix calculus is straightforward. Thus, for $A \in \mathcal{M}_d$, the matrix exponential e^A is defined as usual by Taylor's series. Recall that e^A has the same eigenvectors as A, and eigenvalues $e^{\lambda_i(A)}$. The partial order $A \le B$ means $A - B \ge 0$, i.e. $A - B$ is positive semidefinite.

The non-straightforward part is that, in general, $e^{A+B} \neq e^A e^B$. However, Golden-Thompson's inequality (see [3]) states that

$$\operatorname{tr} e^{A+B} \le \operatorname{tr}(e^A e^B)$$

holds for arbitrary $A, B \in \mathcal{M}_d$ (and in fact for arbitrary unitary-invariant norm replacing the trace).

Therefore, for $S_n = X_1 + \cdots + X_n$ and for I_d being the identity on $\mathcal{M}_d$, we have

$$p := \mathbb{P}(S_n \not\le tI_d) = \mathbb{P}(e^{\lambda S_n} \not\le e^{\lambda t I_d}) \le \mathbb{P}(\operatorname{tr} e^{\lambda S_n} > e^{\lambda t}) \le e^{-\lambda t}\mathbb{E}\operatorname{tr}(e^{\lambda S_n}).$$

This estimate is not sharp: $e^{\lambda S_n} \not\le e^{\lambda t I_d}$ means that the biggest eigenvalue of $e^{\lambda S_n}$ exceeds $e^{\lambda t}$, while $\operatorname{tr} e^{\lambda S_n} > e^{\lambda t}$ means that the sum of all d eigenvalues exceeds the same. This will be responsible for the (sometimes inevitable) loss of the $\log d$ factor in Rudelson's selection theorem.

Since $S_n = X_n + S_{n-1}$, we can use Golden-Thomson's inequality to separate the last term from the sum:

$$\mathbb{E}\operatorname{tr}(e^{\lambda S_n}) \le \mathbb{E}\operatorname{tr}(e^{\lambda X_n} e^{\lambda S_{n-1}}).$$

Now, using independence and that $\mathbb{E}$ and trace commute, we continue to write

$$= \mathbb{E}_{n-1}\operatorname{tr}(\mathbb{E}_n e^{\lambda X_n} \cdot e^{\lambda S_{n-1}}) \le \|\mathbb{E}_n e^{\lambda X_n}\| \cdot \mathbb{E}_{n-1}\operatorname{tr}(e^{\lambda S_{n-1}}).$$

Continuing by induction, we arrive (since $\operatorname{tr}(I_d) = d$) to

$$\mathbb{E}\operatorname{tr}(e^{\lambda S_n}) \le d \cdot \prod_{i=1}^n \|\mathbb{E} e^{\lambda X_i}\|.$$

We have proved that

$$\mathbb{P}(S_n \not\le tI_d) \le de^{-\lambda t} \cdot \prod_{i=1}^n \|\mathbb{E} e^{\lambda X_i}\|.$$

Repeating for $-S_n$ and using that $tI_d \le S_n \le tI_d$ is equivalent to $\|S_n\| \le t$, we have shown that

$$\mathbb{P}(\|S_n\| > t) \le 2de^{-\lambda t} \cdot \prod_{i=1}^n \|\mathbb{E} e^{\lambda X_i}\|. \tag{1}$$

Remark. As in the real valued case, full independence is never needed in the above argument. It works out well for martingales.

The main result:

Theorem 1 (Chernoff-type inequality). *Let $X_i \in \mathcal{M}_d$ be independent mean zero random matrices, $\|X_i\| \le 1$ for all i a.s. Let $S_n = X_1 + \cdots + X_n$, $\sigma^2 = \sum_{i=1}^n \|\operatorname{Var} X_i\|$. Then for every $t > 0$ we have*

$$\mathbb{P}(\|S_n\| > t) \le d \cdot \max(e^{-t^2/4\sigma^2}, e^{-t/2}).$$

To prove this theorem, we have to estimate $\|\mathbb{E}e^{\lambda X_i}\|$ in (1), which is easy.

For example, if $X \in \mathcal{M}_d$ and $\|X\| \le 1$, then Taylor series expansion shows that

$$e^Z \le I_d + Z + Z^2.$$

Therefore, we have

Lemma 2. *Let $Z \in \mathcal{M}_d$ be a mean zero random matrix, $\|Z\| \le 1$ a.s. Then*

$$\mathbb{E}e^Z \le e^{\operatorname{Var} Z}.$$

Proof. Using the mean zero assumption, we have

$$\mathbb{E}e^Z \le \mathbb{E}(I_d + Z + Z^2) = I_d + \operatorname{Var}(Z) \le e^{\operatorname{Var} Z}.$$

□

Let $0 < \lambda \le 1$. Therefore, by the Theorem's hypotheses,

$$\|\mathbb{E}e^{\lambda X_i}\| \le \|e^{\lambda^2 \operatorname{Var} X_i}\| = e^{\lambda^2 \|\operatorname{Var} X_i\|}.$$

Hence by (1),

$$\mathbb{P}(\|S\| > t) \le d \cdot e^{-\lambda t + \lambda^2 \sigma^2}.$$

With the optimal choice of $\lambda := \min(t/2\sigma^2, 1)$, the conclusion of Theorem follows.

Problem: does the Theorem hold for σ^2 replaced by $\|\sum_{i=1}^n \operatorname{Var} X_i\|$? If so, this would generalize Pisier-Lust-Piquard's non-commutative Khinchine inequality.

Corollary 3. *Let $X_i \in \mathcal{M}_d$ be independent random matrices, $X_i \ge 0$, $\|X_i\| \le 1$ for all i a.s. Let $S_n = X_1 + \cdots + X_n$, $E = \sum_{i=1}^n \|\mathbb{E}X_i\|$. Then for every $\varepsilon \in (0,1)$ we have*

$$\mathbb{P}(\|S_n - \mathbb{E}S_n\| > \varepsilon E) \le d \cdot e^{-\varepsilon^2 E/4}.$$

Proof. Applying Theorem for $X_i - \mathbb{E}X_i$, we have

$$\mathbb{P}(\|S_n - \mathbb{E}S_n\| > t) \le d \cdot \max(e^{-t^2/4\sigma^2}, e^{-t/2}). \tag{2}$$

Note that $\|X_i\| \le 1$ implies that

$$\operatorname{Var} X_i \le \mathbb{E}X_i^2 \le \mathbb{E}(\|X_i\| X_i) \le \mathbb{E}X_i.$$

Therefore, $\sigma^2 \le E$. Now we use (2) for $t = \varepsilon E$, and note that

$$t^2/4\sigma^2 = \varepsilon^2 E^2/4\sigma^2 \ge \varepsilon^2 E/4.$$

□

Remark. The hypothesis $\|X_i\| \le 1$ can be relaxed throughout to $\|X_i\|_{\psi_1} \le 1$. One just has to be more careful with Taylor series.

2. Applications

Let x be a random vector in isotropic position in $\mathbb{R}^d$, i.e.

$$\mathbb{E} x \otimes x = I_d.$$

Denote $\|x\|_{\psi_1} = M$. Then the random matrix

$$X := M^{-2} x \otimes x$$

satisfies the hypotheses of Corollary (see remark below it). Clearly,

$$\mathbb{E}X = M^{-2} I_d, \quad E = n/M^2, \quad \mathbb{E}S_n = (n/M^2) I_d.$$

Then Corollary gives

$$\mathbb{P}(\|S_n - \mathbb{E}S_n\| > \varepsilon \|\mathbb{E}S\|) \le d \cdot e^{-\varepsilon^2 n/4M^2}.$$

We have thus proved:

Corollary 4. *Let x be a random vector in isotropic position in $\mathbb{R}^d$, such that $M := \|x\|_{\psi_1} < \infty$. Let $x_1, \ldots, x_n$ be independent copies of x. Then for every $\varepsilon \in (0,1)$, we have*

$$\mathbb{P}\Big(\Big\| \frac{1}{n} \sum_{i=1}^n x_i \otimes x_i - I_d \Big\| > \varepsilon \Big) \le d \cdot e^{-\varepsilon^2 n/4M^2}.$$

The probability in the Corollary is smaller than 1 provided that the number of samples is

$$n \gtrsim \varepsilon^{-2} M^2 \log d.$$

This is a version of Rudelson's sampling theorem, where M played the role of $(\mathbb{E}\|x\|^{\log n})^{1/\log n}$.

One can also deduce the main Lemma in [2] from the Corollary in the previous section. Given vectors x_i in $\mathbb{R}^d$, we are interested in the magnitude of

$$\Big\| \sum_{i=1}^N g_i x_i \otimes x_i \Big\|$$

where g_i are independent Gaussians. For normalization, we can assume that

$$\sum_{i=1}^N x_i \otimes x_i = A, \quad \|A\| = 1.$$

Denote

$$M := \max_i \|x_i\|.$$

and consider the random operator

$$X := M^{-2} x_i \otimes x_i \quad \text{with probability} \quad 1/N.$$

As before, let $X_1, X_2, \dots, X_n$ be independent copies of X, and $S = X_1 + \dots + X_n$.

This time, we are going to let $n \to \infty$. By the Central Limit Theorem, the properly scaled sum $S - \mathbb{E}S$ will converge to $\sum_{i=1}^N g_i x_i \otimes x_i$. One then chooses the parameters correctly to produce a version of the main Lemma in [2]. We omit the details.

References

[1] R. Ahlswede, A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Information Theory 48 (2002), 568–579

[2] M. Rudelson, *Random vectors in the isotropic position*

[3] A. Wigderson, D. Xiao, *Derandomzing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications*, Theory of Computing 4 (2008), 53–76