# 4 Concentration Inequalities, Scalar and Matrix Versions

## 4.1 Large Deviation Inequalities

Concentration and large deviations inequalities are among the most useful tools when understanding the performance of some algorithms. In a nutshell they control the probability of a random variable being very far from its expectation.

The simplest such inequality is Markov's inequality:

**Theorem 4.1 (Markov's Inequality)** *Let $X \geq 0$ be a non-negative random variable with $\mathbb{E}[X] < \infty$. Then,*

$$\text{Prob}\{X > t\} \leq \frac{\mathbb{E}[X]}{t}. \tag{31}$$

*Proof.* Let $t > 0$. Define a random variable $Y_t$ as

$$Y_t = \begin{cases} 0 & \text{if } X \leq t \\ t & \text{if } X > t \end{cases}$$

Clearly, $Y_t \leq X$, hence $\mathbb{E}[Y_t] \leq \mathbb{E}[X]$, and

$$t\,\text{Prob}\{X > t\} = \mathbb{E}[Y_t] \leq \mathbb{E}[X],$$

concluding the proof. □

Markov's inequality can be used to obtain many more concentration inequalities. Chebyshev's inequality is a simple inequality that control fluctuations from the mean.

**Theorem 4.2 (Chebyshev's inequality)** *Let $X$ be a random variable with $\mathbb{E}[X^2] < \infty$. Then,*

$$\text{Prob}\{|X - \mathbb{E}X| > t\} \leq \frac{\text{Var}(X)}{t^2}.$$

*Proof.* Apply Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$ to get:

$$\text{Prob}\{|X - \mathbb{E}X| > t\} = \text{Prob}\{(X - \mathbb{E}X)^2 > t^2\} \leq \frac{\mathbb{E}\left[(X - \mathbb{E}X)^2\right]}{t^2} = \frac{\text{Var}(X)}{t^2}.$$

□

### 4.1.1 Sums of independent random variables

In what follows we'll show two useful inequalities involving sums of independent random variables. The intuitive idea is that if we have a sum of independent random variables

$$X = X_1 + \cdots + X_n,$$

where $X_i$ are iid centered random variables, then while the value of $X$ can be of order $\mathcal{O}(n)$ it will very likely be of order $\mathcal{O}(\sqrt{n})$ (note that this is the order of its standard deviation). The inequalities that follow are ways of very precisely controlling the probability of $X$ being larger than $\mathcal{O}(\sqrt{n})$. While we could use, for example, Chebyshev's inequality for this, in the inequalities that follow the probabilities will be exponentially small, rather than quadratic, which will be crucial in many applications to come.

**Theorem 4.3 (Hoeffding's Inequality)** *Let $X_1, X_2, \ldots, X_n$ be independent bounded random variables, i.e., $|X_i| \le a$ and $\mathbb{E}[X_i] = 0$. Then,*

$$\mathrm{Prob}\left\{\left|\sum_{i=1}^n X_i\right| > t\right\} \le 2\exp\left(-\frac{t^2}{2na^2}\right).$$

The inequality implies that fluctuations larger than $\mathcal{O}(\sqrt{n})$ have small probability. For example, for $t = a\sqrt{2n\log n}$ we get that the probability is at most $\frac{2}{n}$.

*Proof.* We first get a probability bound for the event $\sum_{i=1}^n X_i > t$. The proof, again, will follow from Markov. Since we want an exponentially small probability, we use a classical trick that involves exponentiating with any $\lambda > 0$ and then choosing the optimal $\lambda$.

$$\begin{aligned}
\mathrm{Prob}\left\{\sum_{i=1}^n X_i > t\right\} &= \mathrm{Prob}\left\{\sum_{i=1}^n X_i > t\right\} && (32)\\
&= \mathrm{Prob}\left\{e^{\lambda\sum_{i=1}^n X_i} > e^{\lambda t}\right\}\\
&\le \frac{\mathbb{E}[e^{\lambda\sum_{i=1}^n X_i}]}{e^{t\lambda}}\\
&= e^{-t\lambda}\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}], && (33)
\end{aligned}$$

where the penultimate step follows from Markov's inequality and the last equality follows from independence of the $X_i$'s.

We now use the fact that $|X_i| \le a$ to bound $\mathbb{E}[e^{\lambda X_i}]$. Because the function $f(x) = e^{\lambda x}$ is convex,

$$e^{\lambda x} \le \frac{a+x}{2a}e^{\lambda a} + \frac{a-x}{2a}e^{-\lambda a},$$

for all $x \in [-a, a]$.

Since, for all $i$, $\mathbb{E}[X_i] = 0$ we get

$$\mathbb{E}[e^{\lambda X_i}] \le \mathbb{E}\left[\frac{a+X_i}{2a}e^{\lambda a} + \frac{a-X_i}{2a}e^{-\lambda a}\right] \le \frac{1}{2}\left(e^{\lambda a} + e^{-\lambda a}\right) = \cosh(\lambda a)$$

Note that[15]

$$\cosh(x) \le e^{x^2/2}, \quad \text{for all } x \in \mathbb{R}$$

Hence,

$$\mathbb{E}[e^{\lambda X_i}] \le \mathbb{E}[e^{(\lambda X_i)^2/2}] \le e^{(\lambda a)^2/2}.$$

Together with (32), this gives

$$\begin{aligned}
\mathrm{Prob}\left\{\sum_{i=1}^n X_i > t\right\} &\le e^{-t\lambda}\prod_{i=1}^n e^{(\lambda a)^2/2}\\
&= e^{-t\lambda}e^{n(\lambda a)^2/2}
\end{aligned}$$

[15]This follows immediately from the Taylor expansions: $\cosh(x) = \sum_{n=0}^\infty \frac{x^{2n}}{(2n)!}$, $e^{x^2/2} = \sum_{n=0}^\infty \frac{x^{2n}}{2^n n!}$, and $(2n)! \ge 2^n n!$.

This inequality holds for any choice of $\lambda \geq 0$, so we choose the value of $\lambda$ that minimizes

$$\min_{\lambda} \left\{ n\frac{(\lambda a)^2}{2} - t\lambda \right\}$$

Differentiating readily shows that the minimizer is given by

$$\lambda = \frac{t}{na^2},$$

which satisfies $\lambda > 0$. For this choice of $\lambda$,

$$n(\lambda a)^2/2 - t\lambda = \frac{1}{n}\left(\frac{t^2}{2a^2} - \frac{t^2}{a^2}\right) = -\frac{t^2}{2na^2}$$

Thus,

$$\mathrm{Prob}\left\{\sum_{i=1}^{n} X_i > t\right\} \quad \leq \quad e^{-\frac{t^2}{2na^2}}$$

By using the same argument on $\sum_{i=1}^{n}(-X_i)$, and union bounding over the two events we get,

$$\mathrm{Prob}\left\{\left|\sum_{i=1}^{n} X_i\right| > t\right\} \quad \leq \quad 2e^{-\frac{t^2}{2na^2}}$$

□

**Remark 4.4** *Let's say that we have random variables $r_1, \dots, r_n$ i.i.d. distributed as*

$$r_i = \begin{cases} -1 & \text{with probability } p/2 \\ 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p/2. \end{cases}$$

*Then, $\mathbb{E}(r_i) = 0$ and $|r_i| \leq 1$ so Hoeffding's inequality gives:*

$$\mathrm{Prob}\left\{\left|\sum_{i=1}^{n} r_i\right| > t\right\} \leq 2\exp\left(-\frac{t^2}{2n}\right).$$

*Intuitively, the smallest $p$ is the more concentrated $|\sum_{i=1}^{n} r_i|$ should be, however Hoeffding's inequality does not capture this behavior.*

A natural way to quantify this intuition is by noting that the variance of $\sum_{i=1}^{n} r_i$ depends on $p$ as $\mathrm{Var}(r_i) = p$. The inequality that follows, Bernstein's inequality, uses the variance of the summands to improve over Hoeffding's inequality.

The way this is going to be achieved is by strengthening the proof above, more specifically in step (33) we will use the bound on the variance to get a better estimate on $\mathbb{E}[e^{\lambda X_i}]$ essentially by realizing that if $X_i$ is centered, $\mathbb{E}X_i^2 = \sigma^2$, and $|X_i| \leq a$ then, for $k \geq 2$, $\mathbb{E}X_i^k \leq \sigma^2 a^{k-2} = \left(\frac{\sigma^2}{a^2}\right) a^k$.

**Theorem 4.5 (Bernstein's Inequality)** *Let $X_1, X_2, \ldots, X_n$ be independent centered bounded random variables, i.e., $|X_i| \leq a$ and $\mathbb{E}[X_i] = 0$, with variance $\mathbb{E}[X_i^2] = \sigma^2$. Then,*

$$\operatorname{Prob}\left\{\left|\sum_{i=1}^n X_i > t\right|\right\} \leq 2\exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right).$$

**Remark 4.6** *Before proving Bernstein's Inequality, note that on the example of Remark 4.4 we get*

$$\operatorname{Prob}\left\{\left|\sum_{i=1}^n r_i\right| > t\right\} \leq 2\exp\left(-\frac{t^2}{2np + \frac{2}{3}t}\right),$$

*which exhibits a dependence on $p$ and, for small values of $p$ is considerably smaller than what Hoeffding's inequality gives.*

*Proof.*

As before, we will prove

$$\operatorname{Prob}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right),$$

and then union bound with the same result for $-\sum_{i=1}^n X_i$, to prove the Theorem.

For any $\lambda > 0$ we have

$$\begin{aligned}\operatorname{Prob}\left\{\sum_{i=1}^n X_i > t\right\} &= \operatorname{Prob}\{e^{\lambda \sum X_i} > e^{\lambda t}\} \\ &\leq \frac{\mathbb{E}[e^{\lambda \sum X_i}]}{e^{\lambda t}} \\ &= e^{-\lambda t}\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}]\end{aligned}$$

Now comes the source of the improvement over Hoeffding's,

$$\begin{aligned}\mathbb{E}[e^{\lambda X_i}] &= \mathbb{E}\left[1 + \lambda X_i + \sum_{m=2}^\infty \frac{\lambda^m X_i^m}{m!}\right] \\ &\leq 1 + \sum_{m=2}^\infty \frac{\lambda^m a^{m-2}\sigma^2}{m!} \\ &= 1 + \frac{\sigma^2}{a^2}\sum_{m=2}^\infty \frac{(\lambda a)^m}{m!} \\ &= 1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right)\end{aligned}$$

Therefore,

$$\operatorname{Prob}\left\{\sum_{i=1}^n X_i > t\right\} \leq e^{-\lambda t}\left[1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right)\right]^n$$

We will use a few simple inequalities (that can be easily proved with calculus) such as[16] $1 + x \leq e^x$, for all $x \in \mathbb{R}$.

This means that,

$$1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right) \leq e^{\frac{\sigma^2}{a^2}(e^{\lambda a}-1-\lambda a)},$$

which readily implies

$$\text{Prob}\left\{\sum_{i=1}^n X_i > t\right\} \leq e^{-\lambda t} e^{\frac{n\sigma^2}{a^2}(e^{\lambda a}-1-\lambda a)}.$$

As before, we try to find the value of $\lambda > 0$ that minimizes

$$\min_\lambda \left\{-\lambda t + \frac{n\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)\right\}$$

Differentiation gives

$$-t + \frac{n\sigma^2}{a^2}(ae^{\lambda a} - a) = 0$$

which implies that the optimal choice of $\lambda$ is given by

$$\lambda^* = \frac{1}{a}\log\left(1 + \frac{at}{n\sigma^2}\right)$$

If we set

$$u = \frac{at}{n\sigma^2}, \tag{34}$$

then $\lambda^* = \frac{1}{a}\log(1+u)$.

Now, the value of the minimum is given by

$$-\lambda^* t + \frac{n\sigma^2}{a^2}(e^{\lambda^* a} - 1 - \lambda^* a) = -\frac{n\sigma^2}{a^2}\left[(1+u)\log(1+u) - u\right].$$

Which means that,

$$\text{Prob}\left\{\sum_{i=1}^n X_i > t\right\} \quad \leq \quad \exp\left(-\frac{n\sigma^2}{a^2}\left\{(1+u)\log(1+u) - u\right\}\right)$$

The rest of the proof follows by noting that, for every $u > 0$,

$$(1+u)\log(1+u) - u \geq \frac{u}{\frac{2}{u} + \frac{2}{3}}, \tag{35}$$

which implies:

$$\begin{aligned}\text{Prob}\left\{\sum_{i=1}^n X_i > t\right\} \quad &\leq \quad \exp\left(-\frac{n\sigma^2}{a^2}\frac{u}{\frac{2}{u}+\frac{2}{3}}\right) \\ &= \quad \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right).\end{aligned}$$

□

[16] In fact $y = 1 + x$ is a tangent line to the graph of $f(x) = e^x$.

## 4.2 Gaussian Concentration

One of the most important results in concentration of measure is Gaussian concentration, although being a concentration result specific for normally distributed random variables, it will be very useful throughout these lectures. Intuitively it says that if $F : \mathbb{R}^n \to \mathbb{R}$ is a function that is stable in terms of its input then $F(g)$ is very well concentrated around its mean, where $g \in \mathcal{N}(0, I)$. More precisely:

**Theorem 4.7 (Gaussian Concentration)** *Let $X = [X_1, \ldots, X_n]^T$ be a vector with i.i.d. standard Gaussian entries and $F : \mathbb{R}^n \to \mathbb{R}$ a $\sigma$-Lipschitz function (i.e.: $|F(x) - F(y)| \leq \sigma \|x - y\|$, for all $x, y \in \mathbb{R}^n$). Then, for every $t \geq 0$*

$$\mathrm{Prob}\left\{|F(X) - \mathbb{E}F(X)| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

For the sake of simplicity we will show the proof for a slightly weaker bound (in terms of the constant inside the exponent): $\mathrm{Prob}\left\{|F(X) - \mathbb{E}F(X)| \geq t\right\} \leq 2\exp\left(-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right)$. This exposition follows closely the proof of Theorem 2.1.12 in [Tao12] and the original argument is due to Maurey and Pisier. For a proof with the optimal constants see, for example, Theorem 3.25 in these notes [vH14]. We will also assume the function $F$ is smooth — this is actually not a restriction, as a limiting argument can generalize the result from smooth functions to general Lipschitz functions.

*Proof.*

If $F$ is smooth, then it is easy to see that the Lipschitz property implies that, for every $x \in \mathbb{R}^n$, $\|\nabla F(x)\|_2 \leq \sigma$. By subtracting a constant to $F$, we can assume that $\mathbb{E}F(X) = 0$. Also, it is enough to show a one-sided bound

$$\mathrm{Prob}\left\{F(X) - \mathbb{E}F(X) \geq t\right\} \leq \exp\left(-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right),$$

since obtaining the same bound for $-F(X)$ and taking a union bound would gives the result.

We start by using the same idea as in the proof of the large deviation inequalities above; for any $\lambda > 0$, Markov's inequality implies that

$$\begin{aligned}\mathrm{Prob}\left\{F(X) \geq t\right\} &= \mathrm{Prob}\left\{\exp\left(\lambda F(X)\right) \geq \exp\left(\lambda t\right)\right\} \\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]}{\exp\left(\lambda t\right)}\end{aligned}$$

This means we need to upper bound $\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]$ using a bound on $\|\nabla F\|$. The idea is to introduce a random independent copy $Y$ of $X$. Since $\exp\left(\lambda \cdot\right)$ is convex, Jensen's inequality implies that

$$\mathbb{E}\left[\exp\left(-\lambda F(Y)\right)\right] \geq \exp\left(-\mathbb{E}\lambda F(Y)\right) = \exp(0) = 1.$$

Hence, since $X$ and $Y$ are independent,

$$\mathbb{E}\left[\exp\left(\lambda\left[F(X) - F(Y)\right]\right)\right] = \mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]\mathbb{E}\left[\exp\left(-\lambda F(Y)\right)\right] \geq \mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]$$

Now we use the Fundamental Theorem of Calculus in a circular arc from $X$ to $Y$:

$$F(X) - F(Y) = \int_0^{\frac{\pi}{2}} \frac{\partial}{\partial\theta} F\left(Y\cos\theta + X\sin\theta\right) d\theta.$$

The advantage of using the circular arc is that, for any $\theta$, $X_\theta := Y\cos\theta + X\sin\theta$ is another random variable with the same distribution. Also, its derivative with respect to $\theta$, $X'_\theta = -Y\sin\theta + X\cos\theta$ also is. Moreover, $X_\theta$ and $X'_\theta$ are independent. In fact, note that

$$\mathbb{E}\left[X_\theta {X'_\theta}^T\right] = \mathbb{E}\left[Y\cos\theta + X\sin\theta\right]\left[-Y\sin\theta + X\cos\theta\right]^T = 0.$$

We use Jensen's again (with respect to the integral now) to get:

$$\begin{aligned}\exp\left(\lambda\left[F(X) - F(Y)\right]\right) &= \exp\left(\lambda\frac{\pi}{2}\frac{1}{\pi/2}\int_0^{\pi/2}\frac{\partial}{\partial\theta}F\left(X_\theta\right)d\theta\right)\\ &\leq \frac{1}{\pi/2}\int_0^{\pi/2}\exp\left(\lambda\frac{\pi}{2}\frac{\partial}{\partial\theta}F\left(X_\theta\right)\right)d\theta\end{aligned}$$

Using the chain rule,

$$\exp\left(\lambda\left[F(X) - F(Y)\right]\right) \leq \frac{2}{\pi}\int_0^{\pi/2}\exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X'_\theta\right)d\theta,$$

and taking expectations

$$\mathbb{E}\exp\left(\lambda\left[F(X) - F(Y)\right]\right) \leq \frac{2}{\pi}\int_0^{\pi/2}\mathbb{E}\exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X'_\theta\right)d\theta,$$

If we condition on $X_\theta$, since $\left\|\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\right\| \leq \lambda\frac{\pi}{2}\sigma$, $\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X'_\theta$ is a gaussian random variable with variance at most $\left(\lambda\frac{\pi}{2}\sigma\right)^2$. This directly implies that, for every value of $X_\theta$

$$\mathbb{E}_{X'_\theta}\exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X'_\theta\right) \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2\right]$$

Taking expectation now in $X_\theta$, and putting everything together, gives

$$\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right] \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2\right],$$

which means that

$$\text{Prob}\left\{F(X) \geq t\right\} \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2 - \lambda t\right],$$

Optimizing for $\lambda$ gives $\lambda^* = \left(\frac{2}{\pi}\right)^2\frac{t}{\sigma^2}$, which gives

$$\text{Prob}\left\{F(X) \geq t\right\} \leq \exp\left[-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right].$$

□

### 4.2.1 Spectral norm of a Wigner Matrix

We give an illustrative example of the utility of Gaussian concentration. Let $W \in \mathbb{R}^{n\times n}$ be a standard Gaussian Wigner matrix, a symmetric matrix with (otherwise) independent gaussian entries, the off-diagonal entries have unit variance and the diagonal entries have variance 2. $\|W\|$ depends on $\frac{n(n+1)}{2}$ independent (standard) gaussian random variables and it is easy to see that it is a $\sqrt{2}$-Lipschitz function of these variables, since

$$\left| \|W^{(1)}\| - \|W^{(2)}\| \right| \leq \left\| W^{(1)} - W^{(2)} \right\| \leq \left\| W^{(1)} - W^{(2)} \right\|_F.$$

The symmetry of the matrix and the variance 2 of the diagonal entries are responsible for an extra factor of $\sqrt{2}$.

Using Gaussian Concentration (Theorem 4.7) we immediately get

$$\operatorname{Prob}\left\{ \|W\| \geq \mathbb{E}\|W\| + t \right\} \leq \exp\left(-\frac{t^2}{4}\right).$$

Since[17] $\mathbb{E}\|W\| \leq 2\sqrt{n}$ we get

**Proposition 4.8** *Let $W \in \mathbb{R}^{n\times n}$ be a standard Gaussian Wigner matrix, a symmetric matrix with (otherwise) independent gaussian entries, the off-diagonal entries have unit variance and the diagonal entries have variance 2. Then,*

$$\operatorname{Prob}\left\{ \|W\| \geq 2\sqrt{n} + t \right\} \leq \exp\left(-\frac{t^2}{4}\right).$$

Note that this gives an extremely precise control of the fluctuations of $\|W\|$. In fact, for $t = 2\sqrt{\log n}$ this gives

$$\operatorname{Prob}\left\{ \|W\| \geq 2\sqrt{n} + 2\sqrt{\log n} \right\} \leq \exp\left(-\frac{4\log n}{4}\right) = \frac{1}{n}.$$

### 4.2.2 Talagrand's concentration inequality

A remarkable result by Talagrand [Tal95], Talangrad's concentration inequality, provides an analogue of Gaussian concentration to bounded random variables.

**Theorem 4.9 (Talangrand concentration inequality, Theorem 2.1.13 [Tao12])** *Let $K > 0$, and let $X_1, \ldots, X_n$ be independent bounded random variables, $|X_i| \leq K$ for all $1 \leq i \leq n$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a $\sigma$-Lipschitz and convex function. Then, for any $t \geq 0$,*

$$\operatorname{Prob}\left\{ |F(X) - \mathbb{E}\left[F(X)\right]| \geq tK \right\} \leq c_1 \exp\left(-c_2 \frac{t^2}{\sigma^2}\right),$$

*for positive constants $c_1$, and $c_2$.*

Other useful similar inequalities (with explicit constants) are available in [Mas00].

[17] It is an excellent exercise to prove $\mathbb{E}\|W\| \leq 2\sqrt{n}$ using Slepian's inequality.

### 4.3 Other useful large deviation inequalities

This Section contains, without proof, some scalar large deviation inequalities that I have found useful.

#### 4.3.1 Additive Chernoff Bound

The additive Chernoff bound, also known as Chernoff-Hoeffding theorem concerns Bernoulli random variables.

**Theorem 4.10** *Given $0 < p < 1$ and $X_1, \dots, X_n$ i.i.d. random variables distributed as* Bernoulli($p$) *random variable (meaning that it is 1 with probability $p$ and 0 with probability $1-p$), then, for any $\varepsilon > 0$:*

- $$\text{Prob}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\varepsilon\right\} \leq \left[\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right]^n$$
- $$\text{Prob}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\varepsilon\right\} \leq \left[\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right]^n$$

#### 4.3.2 Multiplicative Chernoff Bound

There is also a multiplicative version (see, for example Lemma 2.3.3. in [Dur06]), which is particularly useful.

**Theorem 4.11** *Let $X_1, \dots, X_n$ be independent random variables taking values is $\{0,1\}$ (meaning they are Bernoulli distributed but not necessarily identically distributed). Let $\mu = \mathbb{E}\sum_{i=1}^{n} X_i$, then, for any $\delta > 0$:*

- $$\text{Prob}\left\{X > (1+\delta)\mu\right\} < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$$
- $$\text{Prob}\left\{X < (1-\delta)\mu\right\} < \left[\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right]^{\mu}$$

#### 4.3.3 Deviation bounds on $\chi_2$ variables

A particularly useful deviation inequality is Lemma 1 in Laurent and Massart [LM00]:

**Theorem 4.12 (Lemma 1 in Laurent and Massart [LM00])** *Let $X_1, \dots, X_n$ be i.i.d. standard gaussian random variables ($\mathcal{N}(0,1)$), and $a_1, \dots, a_n$ non-negative numbers. Let*

$$Z = \sum_{k=1}^{n} a_k \left(X_k^2 - 1\right).$$

*The following inequalities hold for any $t > 0$:*

- $\text{Prob}\left\{Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x\right\} \leq \exp(-x)$,

- Prob $\{Z \leq -2\|a\|_2\sqrt{x}\} \leq \exp(-x)$,

*where* $\|a\|_2^2 = \sum_{k=1}^n a_k^2$ *and* $\|a\|_\infty = \max_{1\leq k\leq n} |a_k|$.

Note that if $a_k = 1$, for all $k$, then $Z$ is a $\chi_2$ with $n$ degrees of freedom, so this theorem immediately gives a deviation inequality for $\chi_2$ random variables.

## 4.4 Matrix Concentration

In many important applications, some of which we will see in the proceeding lectures, one needs to use a matrix version of the inequalities above.

Given $\{X_k\}_{k=1}^n$ independent random symmetric $d \times d$ matrices one is interested in deviation inequalities for

$$\lambda_{\max}\left(\sum_{k=1}^n X_k\right).$$

For example, a very useful adaptation of Bernstein's inequality exists for this setting.

**Theorem 4.13 (Theorem 1.4 in [Tro12])** *Let* $\{X_k\}_{k=1}^n$ *be a sequence of independent random symmetric* $d \times d$ *matrices. Assume that each* $X_k$ *satisfies:*

$$\mathbb{E}X_k = 0 \text{ and } \lambda_{\max}(X_k) \leq R \text{ almost surely.}$$

*Then, for all* $t \geq 0$,

$$\text{Prob}\left\{\lambda_{\max}\left(\sum_{k=1}^n X_k\right) \geq t\right\} \leq d \cdot \exp\left(\frac{-t^2}{2\sigma^2 + \frac{2}{3}Rt}\right) \text{ where } \sigma^2 = \left\|\sum_{k=1}^n \mathbb{E}\left(X_k^2\right)\right\|.$$

Note that $\|A\|$ denotes the spectral norm of $A$.

In what follows we will state and prove various matrix concentration results, somewhat similar to Theorem 4.13. Motivated by the derivation of Proposition 4.8, that allowed us to easily transform bounds on the expected spectral norm of a random matrix into tail bounds, we will mostly focus on bounding the expected spectral norm. Tropp's monograph [Tro15b] is a nice introduction to matrix concentration and includes a proof of Theorem 4.13 as well as many other useful inequalities.

A particularly important inequality of this type is for gaussian series, it is intimately related to the non-commutative Khintchine inequality [Pis03], and for that reason we will often refer to it as Non-commutative Khintchine (see, for example, (4.9) in [Tro12]).

**Theorem 4.14 (Non-commutative Khintchine (NCK))** *Let* $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ *be symmetric matrices and* $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ *i.i.d., then:*

$$\mathbb{E}\left\|\sum_{k=1}^n g_k A_k\right\| \leq \left(2 + 2\log(2d)\right)^{\frac{1}{2}}\sigma,$$

*where*

$$\sigma^2 = \left\|\sum_{k=1}^n A_k^2\right\|^2. \tag{36}$$

Note that, akin to Proposition 4.8, we can also use Gaussian Concentration to get a tail bound on $\|\sum_{k=1}^{n} g_k A_k\|$. We consider the function

$$F : \mathbb{R}^n \to \left\| \sum_{k=1}^{n} g_k A_k \right\|.$$

We now estimate its Lipschitz constant; let $g, h \in \mathbb{R}^n$ then

$$\begin{aligned}
\left| \left\| \sum_{k=1}^{n} g_k A_k \right\| - \left\| \sum_{k=1}^{n} h_k A_k \right\| \right| &\leq \left\| \left( \sum_{k=1}^{n} g_k A_k \right) - \left( \sum_{k=1}^{n} h_k A_k \right) \right\| \\
&= \left\| \sum_{k=1}^{n} (g_k - h_k) A_k \right\| \\
&= \max_{v:\, \|v\|=1} v^T \left( \sum_{k=1}^{n} (g_k - h_k) A_k \right) v \\
&= \max_{v:\, \|v\|=1} \sum_{k=1}^{n} (g_k - h_k) \left( v^T A_k v \right) \\
&\leq \max_{v:\, \|v\|=1} \sqrt{\sum_{k=1}^{n} (g_k - h_k)^2} \sqrt{\sum_{k=1}^{n} \left( v^T A_k v \right)^2} \\
&= \sqrt{\max_{v:\, \|v\|=1} \sum_{k=1}^{n} \left( v^T A_k v \right)^2} \|g - h\|_2,
\end{aligned}$$

where the first inequality made use of the triangular inequality and the last one of the Cauchy-Schwarz inequality.

This motivates us to define a new parameter, the weak variance $\sigma_*$.

**Definition 4.15 (Weak Variance (see, for example, [Tro15b]))** *Given $A_1, \ldots, A_n \in \mathbb{R}^{d \times d}$ symmetric matrices. We define the weak variance parameter as*

$$\sigma_*^2 = \max_{v:\, \|v\|=1} \sum_{k=1}^{n} \left( v^T A_k v \right)^2.$$

This means that, using Gaussian concentration (and setting $t = u\sigma_*$), we have

$$\text{Prob}\left\{ \left\| \sum_{k=1}^{n} g_k A_k \right\| \geq \left( 2 + 2\log(2d) \right)^{\frac{1}{2}} \sigma + u\sigma_* \right\} \leq \exp\left( -\frac{1}{2} u^2 \right). \tag{37}$$

This means that although the expected value of $\|\sum_{k=1}^{n} g_k A_k\|$ is controlled by the parameter $\sigma$, its fluctuations seem to be controlled by $\sigma_*$. We compare the two quantities in the following Proposition.

**Proposition 4.16** *Given $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ symmetric matrices, recall that*

$$\sigma = \sqrt{\left\|\sum_{k=1}^{n} A_k^2\right\|^2} \text{ and } \sigma_* = \sqrt{\max_{v:\,\|v\|=1} \sum_{k=1}^{n} (v^T A_k v)^2}.$$

*We have*

$$\sigma_* \leq \sigma.$$

*Proof.* Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sigma_*^2 &= \max_{v:\,\|v\|=1} \sum_{k=1}^{n} \left(v^T A_k v\right)^2 \\
&= \max_{v:\,\|v\|=1} \sum_{k=1}^{n} \left(v^T [A_k v]\right)^2 \\
&\leq \max_{v:\,\|v\|=1} \sum_{k=1}^{n} (\|v\|\|A_k v\|)^2 \\
&= \max_{v:\,\|v\|=1} \sum_{k=1}^{n} \|A_k v\|^2 \\
&= \max_{v:\,\|v\|=1} \sum_{k=1}^{n} v^T A_k^2 v \\
&= \left\|\sum_{k=1}^{n} A_k^2\right\| \\
&= \sigma^2.
\end{aligned}$$

□

## 4.5 Optimality of matrix concentration result for gaussian series

The following simple calculation is suggestive that the parameter $\sigma$ in Theorem 4.14 is indeed the correct parameter to understand $\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\|$.

$$\begin{aligned}
\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\|^2 &= \mathbb{E}\left\|\left(\sum_{k=1}^{n} g_k A_k\right)^2\right\| = \mathbb{E} \max_{v:\ \|v\|=1} v^T \left(\sum_{k=1}^{n} g_k A_k\right)^2 v \\
&\geq \max_{v:\ \|v\|=1} \mathbb{E} v^T \left(\sum_{k=1}^{n} g_k A_k\right)^2 v = \max_{v:\ \|v\|=1} v^T \left(\sum_{k=1}^{n} A_k^2\right) v = \sigma^2 \qquad (38)
\end{aligned}$$

But a natural question is whether the logarithmic term is needed. Motivated by this question we'll explore a couple of examples.

**Example 4.17** *We can write a $d \times d$ Wigner matrix $W$ as a gaussian series, by taking $A_{ij}$ for $i \le j$ defined as*

$$A_{ij} = e_i e_j^T + e_j e_i^T,$$

*if $i \ne j$, and*

$$A_{ii} = \sqrt{2} e_i e_i^T.$$

*It is not difficult to see that, in this case, $\sum_{i \le j} A_{ij}^2 = (d+1) I_{d \times d}$, meaning that $\sigma = \sqrt{d+1}$. This means that Theorem 4.14 gives us*

$$\mathbb{E}\|W\| \lesssim \sqrt{d \log d},$$

*however, we know that $\mathbb{E}\|W\| \asymp \sqrt{d}$, meaning that the bound given by NCK (Theorem 4.14) is, in this case, suboptimal by a logarithmic factor.*[18]

The next example will show that the logarithmic factor is in fact needed in some examples

**Example 4.18** *Consider $A_k = e_k e_k^T \in \mathbb{R}^{d \times d}$ for $k = 1, \dots, d$. The matrix $\sum_{k=1}^n g_k A_k$ corresponds to a diagonal matrix with independent standard gaussian random variables as diagonal entries, and so it's spectral norm is given by $\max_k |g_k|$. It is known that $\max_{1 \le k \le d} |g_k| \asymp \sqrt{\log d}$. On the other hand, a direct calculation shows that $\sigma = 1$. This shows that the logarithmic factor cannot, in general, be removed.*

This motivates the question of trying to understand when is it that the extra dimensional factor is needed. For both these examples, the resulting matrix $X = \sum_{k=1}^n g_k A_k$ has independent entries (except for the fact that it is symmetric). The case of independent entries [RS13, Seg00, Lat05, BvH15] is now somewhat understood:

**Theorem 4.19 ([BvH15])** *If $X$ is a $d \times d$ random symmetric matrix with gaussian independent entries (except for the symmetry constraint) whose entry $i, j$ has variance $b_{ij}^2$ then*

$$\mathbb{E}\|X\| \lesssim \sqrt{\max_{1 \le i \le d} \sum_{j=1}^{d} b_{ij}^2} + \max_{ij} |b_{ij}| \sqrt{\log d}.$$

**Remark 4.20** *$X$ in the theorem above can be written in terms of a Gaussian series by taking*

$$A_{ij} = b_{ij} \left( e_i e_j^T + e_j e_i^T \right),$$

*for $i < j$ and $A_{ii} = b_{ii} e_i e_i^T$. One can then compute $\sigma$ and $\sigma_*$:*

$$\sigma^2 = \max_{1 \le i \le d} \sum_{j=1}^{d} b_{ij}^2 \text{ and } \sigma_*^2 \asymp b_{ij}^2.$$

*This means that, when the random matrix in NCK (Theorem 4.14) has negative entries (modulo symmetry) then*

$$\mathbb{E}\|X\| \lesssim \sigma + \sqrt{\log d}\,\sigma_*. \tag{39}$$

[18] By $a \asymp b$ we mean $a \lesssim b$ and $a \gtrsim b$.

Theorem 4.19 together with a recent improvement of Theorem 4.14 by Tropp [Tro15c][19] motivate the bold possibility of (39) holding in more generality.

**Conjecture 4.21** *Let $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ be symmetric matrices and $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ i.i.d., then:*

$$\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\| \lesssim \sigma + (\log d)^{\frac{1}{2}}\,\sigma_*,$$

While it may very will be that this Conjecture 4.21 is false, no counter example is known, up to date.

**Open Problem 4.1 (Improvement on Non-Commutative Khintchine Inequality)** *Prove or disprove Conjecture 4.21.*

I would also be pretty excited to see interesting examples that satisfy the bound in Conjecture 4.21 while such a bound would not trivially follow from Theorems 4.14 or 4.19.

### 4.5.1 An interesting observation regarding random matrices with independent matrices

For the independent entries setting, Theorem 4.19 is tight (up to constants) for a wide range of variance profiles $\left\{b_{ij}^2\right\}_{i\le j}$ – the details are available as Corollary 3.15 in [BvH15]; the basic idea is that if the largest variance is comparable to the variance of a sufficient number of entries, then the bound in Theorem 4.19 is tight up to constants.

However, the situation is not as well understood when the variance profiles $\left\{b_{ij}^2\right\}_{i\le j}$ are arbitrary. Since the spectral norm of a matrix is always at least the $\ell_2$ norm of a row, the following lower bound holds (for $X$ a symmetric random matrix with independent gaussian entries):

$$\mathbb{E}\|X\| \geq \mathbb{E}\max_k \|Xe_k\|_2.$$

Observations in papers of Latała [Lat05] and Riemer and Schutt [RS13], together with the results in [BvH15], motivate the conjecture that this lower bound is always tight (up to constants).

**Open Problem 4.2 (Latała-Riemer-Schutt)** *Given $X$ a symmetric random matrix with independent gaussian entries, is the following true?*

$$\mathbb{E}\|X\| \lesssim \mathbb{E}\max_k \|Xe_k\|_2.$$

The results in [BvH15] answer this in the positive for a large range of variance profiles, but not in full generality. Recently, van Handel [vH15] proved this conjecture in the positive with an extra factor of $\sqrt{\log\log d}$. More precisely, that

$$\mathbb{E}\|X\| \lesssim \sqrt{\log\log d}\,\mathbb{E}\max_k \|Xe_k\|_2,$$

where $d$ is the number of rows (and columns) of $X$.

[19] We briefly discuss this improvement in Remark 4.32

### 4.6 A matrix concentration inequality for Rademacher Series

In what follows, we closely follow [Tro15a] and present an elementary proof of a few useful matrix concentration inequalities. We start with a Master Theorem of sorts for Rademacher series (the Rademacher analogue of Theorem 4.14)

**Theorem 4.22** *Let $H_1, \dots, H_n \in \mathbb{R}^{d\times d}$ be symmetric matrices and $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. Rademacher random variables (meaning $= +1$ with probability $1/2$ and $= -1$ with probability $1/2$), then:*

$$\mathbb{E}\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \leq \Big(1 + 2\lceil \log(d)\rceil\Big)^{\frac{1}{2}}\sigma,$$

*where*

$$\sigma^2 = \left\|\sum_{k=1}^n H_k^2\right\|^2. \tag{40}$$

Before proving this theorem, we take first a small detour in discrepancy theory followed by derivations, using this theorem, of a couple of useful matrix concentration inequalities.

#### 4.6.1 A small detour on discrepancy theory

The following conjecture appears in a nice blog post of Raghu Meka [Mek14].

**Conjecture 4.23** *[Matrix Six-Deviations Suffice] There exists a universal constant $C$ such that, for any choice of $n$ symmetric matrices $H_1, \dots, H_n \in \mathbb{R}^{n\times n}$ satisfying $\|H_k\| \leq 1$ (for all $k = 1, \dots, n$), there exists $\varepsilon_1, \dots, \varepsilon_n \in \{\pm 1\}$ such that*

$$\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \leq C\sqrt{n}.$$

**Open Problem 4.3** *Prove or disprove Conjecture 4.23.*

Note that, when the matrices $H_k$ are diagonal, this problem corresponds to Spencer's Six Standard Deviations Suffice Theorem [Spe85].

**Remark 4.24** *Also, using Theorem 4.22, it is easy to show that if one picks $\varepsilon_i$ as i.i.d. Rademacher random variables, then with positive probability (via the probabilistic method) the inequality will be satisfied with an extra $\sqrt{\log n}$ term. In fact one has*

$$\mathbb{E}\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \lesssim \sqrt{\log n}\sqrt{\left\|\sum_{k=1}^n H_k^2\right\|} \leq \sqrt{\log n}\sqrt{\sum_{k=1}^n \|H_k\|^2} \leq \sqrt{\log n}\sqrt{n}.$$

**Remark 4.25** *Remark 4.24 motivates asking whether Conjecture 4.23 can be strengthened to ask for $\varepsilon_1, \dots, \varepsilon_n$ such that*

$$\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \lesssim \left\|\sum_{k=1}^n H_k^2\right\|^{\frac{1}{2}}. \tag{41}$$

### 4.6.2 Back to matrix concentration

Using Theorem 4.22, we'll prove the following Theorem.

**Theorem 4.26** *Let $T_1, \ldots, T_n \in \mathbb{R}^{d\times d}$ be random independent positive semidefinite matrices, then*

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left[\left\|\sum_{i=1}^n \mathbb{E}T_i\right\|^{\frac{1}{2}} + \sqrt{C(d)}\left(\mathbb{E}\max_i \|T_i\|\right)^{\frac{1}{2}}\right]^2,$$

*where*

$$C(d) := 4 + 8\lceil \log d \rceil. \tag{42}$$

A key step in the proof of Theorem 4.26 is an idea that is extremely useful in Probability, the trick of symmetrization. For this reason we isolate it in a lemma.

**Lemma 4.27 (Symmetrization)** *Let $T_1, \ldots, T_n$ be independent random matrices (note that they don't necessarily need to be positive semidefinite, for the sake of this lemma) and $\varepsilon_1, \ldots, \varepsilon_n$ random i.i.d. Rademacher random variables (independent also from the matrices). Then*

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left\|\sum_{i=1}^n \mathbb{E}T_i\right\| + 2\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\|$$

*Proof.* Triangular inequality gives

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left\|\sum_{i=1}^n \mathbb{E}T_i\right\| + \mathbb{E}\left\|\sum_{i=1}^n (T_i - \mathbb{E}T_i)\right\|.$$

Let us now introduce, for each $i$, a random matrix $T_i'$ identically distributed to $T_i$ and independent (all $2n$ matrices are independent). Then

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^n (T_i - \mathbb{E}T_i)\right\| &= \mathbb{E}_T\left\|\sum_{i=1}^n \left(T_i - \mathbb{E}T_i - \mathbb{E}_{T_i'}\left[T_i' - \mathbb{E}_{T_i'}T_i'\right]\right)\right\| \\
&= \mathbb{E}_T\left\|\mathbb{E}_{T'}\sum_{i=1}^n \left(T_i - T_i'\right)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^n \left(T_i - T_i'\right)\right\|,
\end{aligned}$$

where we use the notation $\mathbb{E}_a$ to mean that the expectation is taken with respect to the variable $a$ and the last step follows from Jensen's inequality with respect to $\mathbb{E}_{T'}$.

Since $T_i - T_i'$ is a symmetric random variable, it is identically distributed to $\varepsilon_i\left(T_i - T_i'\right)$ which gives

$$\mathbb{E}\left\|\sum_{i=1}^n \left(T_i - T_i'\right)\right\| = \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i\left(T_i - T_i'\right)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\| + \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i'\right\| = 2\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\|,$$

concluding the proof. □

*Proof.* [of Theorem 4.26]

Using Lemma 4.27 and Theorem 4.22 we get

$$\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\| \leq \left\|\sum_{i=1}^{n} \mathbb{E}T_i\right\| + \sqrt{C(d)}\mathbb{E}\left\|\sum_{i=1}^{n} T_i^2\right\|^{\frac{1}{2}}$$

The trick now is to make a term like the one in the LHS appear in the RHS. For that we start by noting (you can see Fact 2.3 in [Tro15a] for an elementary proof) that, since $T_i \succeq 0$,

$$\left\|\sum_{i=1}^{n} T_i^2\right\| \leq \max_i \|T_i\| \left\|\sum_{i=1}^{n} T_i\right\|.$$

This means that

$$\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\| \leq \left\|\sum_{i=1}^{n} \mathbb{E}T_i\right\| + \sqrt{C(d)}\mathbb{E}\left[\left(\max_i \|T_i\|\right)^{\frac{1}{2}} \left\|\sum_{i=1}^{n} T_i\right\|^{\frac{1}{2}}\right].$$

Further applying the Cauchy-Schwarz inequality for $\mathbb{E}$ gives,

$$\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\| \leq \left\|\sum_{i=1}^{n} \mathbb{E}T_i\right\| + \sqrt{C(d)}\left(\mathbb{E}\max_i \|T_i\|\right)^{\frac{1}{2}} \left(\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\|\right)^{\frac{1}{2}},$$

Now that the term $\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\|$ appears in the RHS, the proof can be finished with a simple application of the quadratic formula (see Section 6.1. in [Tro15a] for details).

□

We now show an inequality for general symmetric matrices

**Theorem 4.28** *Let $Y_1, \ldots, Y_n \in \mathbb{R}^{d\times d}$ be random independent positive semidefinite matrices, then*

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i\right\| \leq \sqrt{C(d)}\sigma + C(d)L,$$

*where,*

$$\sigma^2 = \left\|\sum_{i=1}^{n} \mathbb{E}Y_i^2\right\| \text{ and } L^2 = \mathbb{E}\max_i \|Y_i\|^2 \tag{43}$$

*and, as in* (42),

$$C(d) := 4 + 8\lceil \log d \rceil.$$

*Proof.*

Using Symmetrization (Lemma 4.27) and Theorem 4.22, we get

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i\right\| \leq 2\mathbb{E}_Y\left[\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{n} \varepsilon_i Y_i\right\|\right] \leq \sqrt{C(d)}\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|^{\frac{1}{2}}.$$

Jensen's inequality gives

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|^{\frac{1}{2}} \leq \left(\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|\right)^{\frac{1}{2}},$$

and the proof can be concluded by noting that $Y_i^2 \succeq 0$ and using Theorem 4.26. □

**Remark 4.29 (The rectangular case)** *One can extend Theorem 4.28 to general rectangular matrices $S_1, \ldots, S_n \in \mathbb{R}^{d_1 \times d_2}$ by setting*

$$Y_i = \begin{bmatrix} 0 & S_i \\ S_i^T & 0 \end{bmatrix},$$

*and noting that*

$$\|Y_i^2\| = \left\|\begin{bmatrix} 0 & S_i \\ S_i^T & 0 \end{bmatrix}^2\right\| = \left\|\begin{bmatrix} S_i S_i^T & 0 \\ 0 & S_i^T S_i \end{bmatrix}\right\| = \max\left\{\left\|S_i^T S_i\right\|, \left\|S_i S_i^T\right\|\right\}.$$

*We defer the details to [Tro15a]*

In order to prove Theorem 4.22, we will use an AM-GM like inequality for matrices for which, unlike the one on Open Problem 0.2. in [Ban15d], an elementary proof is known.

**Lemma 4.30** *Given symmetric matrices $H, W, Y \in \mathbb{R}^{d \times d}$ and non-negative integers $r, q$ satisfying $q \leq 2r$,*

$$\operatorname{Tr}\left[HW^qHY^{2r-q}\right] + \operatorname{Tr}\left[HW^{2r-q}HY^q\right] \leq \operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right],$$

*and summing over $q$ gives*

$$\sum_{q=0}^{2r} \operatorname{Tr}\left[HW^qHY^{2r-q}\right] \leq \left(\frac{2r+1}{2}\right) \operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right]$$

We refer to Fact 2.4 in [Tro15a] for an elementary proof but note that it is a matrix analogue to the inequality,

$$\mu^\theta \lambda^{1-\theta} + \mu^{1-\theta}\lambda^\theta \leq \lambda + \theta$$

for $\mu, \lambda \geq 0$ and $0 \leq \theta \leq 1$, which can be easily shown by adding two AM-GM inequalities

$$\mu^\theta \lambda^{1-\theta} \leq \theta\mu + (1-\theta)\lambda \text{ and } \mu^{1-\theta}\lambda^\theta \leq (1-\theta)\mu + \theta\lambda.$$

*Proof.* [of Theorem 4.22]

Let $X = \sum_{k=1}^{n} \varepsilon_k H_k$, then for any positive integer $p$,

$$\mathbb{E}\|X\| \leq \left(\mathbb{E}\|X\|^{2p}\right)^{\frac{1}{2p}} = \left(\mathbb{E}\|X^{2p}\|\right)^{\frac{1}{2p}} \leq \left(\mathbb{E}\operatorname{Tr} X^{2p}\right)^{\frac{1}{2p}},$$

where the first inequality follows from Jensen's inequality and the last from $X^{2p} \succeq 0$ and the observation that the trace of a positive semidefinite matrix is at least its spectral norm. In the sequel, we upper bound $\mathbb{E}\operatorname{Tr} X^{2p}$. We introduce $X_{+i}$ and $X_{-i}$ as $X$ conditioned on $\varepsilon_i$ being, respectively $+1$ or $-1$. More precisely

$$X_{+i} = H_i + \sum_{j \neq i} \varepsilon_j H_j \text{ and } X_{-i} = -H_i + \sum_{j \neq i} \varepsilon_j H_j.$$

Then, we have

$$\mathbb{E}\operatorname{Tr} X^{2p} = \mathbb{E}\operatorname{Tr}\left[X X^{2p-1}\right] = \mathbb{E}\sum_{i=1}^{n} \operatorname{Tr} \varepsilon_i H_i X^{2p-1}.$$

Note that $\mathbb{E}_{\varepsilon_i} \operatorname{Tr}\left[\varepsilon_i H_i X^{2p-1}\right] = \frac{1}{2}\operatorname{Tr}\left[H_i\left(X_{+i}^{2p-1} - X_{-i}^{2p-1}\right)\right]$, this means that

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^{n} \mathbb{E}\frac{1}{2}\operatorname{Tr}\left[H_i\left(X_{+i}^{2p-1} - X_{-i}^{2p-1}\right)\right],$$

where the expectation can be taken over $\varepsilon_j$ for $j \neq i$.

Now we rewrite $X_{+i}^{2p-1} - X_{-i}^{2p-1}$ as a telescopic sum:

$$X_{+i}^{2p-1} - X_{-i}^{2p-1} = \sum_{q=0}^{2p-2} X_{+i}^{q}\left(X_{+i} - X_{-i}\right) X_{-i}^{2p-2-q}.$$

Which gives

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^{n}\sum_{q=0}^{2p-2} \mathbb{E}\frac{1}{2}\operatorname{Tr}\left[H_i X_{+i}^{q}\left(X_{+i} - X_{-i}\right) X_{-i}^{2p-2-q}\right].$$

Since $X_{+i} - X_{-i} = 2H_i$ we get

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^{n}\sum_{q=0}^{2p-2} \mathbb{E}\operatorname{Tr}\left[H_i X_{+i}^{q} H_i X_{-i}^{2p-2-q}\right]. \tag{44}$$

We now make use of Lemma 4.30 to get[20] to get

$$\mathbb{E}\operatorname{Tr} X^{2p} \leq \sum_{i=1}^{n} \frac{2p-1}{2}\mathbb{E}\operatorname{Tr}\left[H_i^2\left(X_{+i}^{2p-2} + X_{-i}^{2p-2}\right)\right]. \tag{45}$$

[20] See Remark 4.32 regarding the suboptimality of this step.

Hence,

$$\begin{aligned}
\sum_{i=1}^{n}\frac{2p-1}{2}\mathbb{E}\operatorname{Tr}\left[H_i^2\left(X_{+i}^{2p-2}+X_{-i}^{2p-2}\right)\right] &= (2p-1)\sum_{i=1}^{n}\mathbb{E}\operatorname{Tr}\left[H_i^2\frac{\left(X_{+i}^{2p-2}+X_{-i}^{2p-2}\right)}{2}\right]\\
&= (2p-1)\sum_{i=1}^{n}\mathbb{E}\operatorname{Tr}\left[H_i^2\mathbb{E}_{\varepsilon_i}\left[X^{2p-2}\right]\right]\\
&= (2p-1)\sum_{i=1}^{n}\mathbb{E}\operatorname{Tr}\left[H_i^2X^{2p-2}\right]\\
&= (2p-1)\mathbb{E}\operatorname{Tr}\left[\left(\sum_{i=1}^{n}H_i^2\right)X^{2p-2}\right]
\end{aligned}$$

Since $X^{2p-2}\succeq 0$ we have

$$\operatorname{Tr}\left[\left(\sum_{i=1}^{n}H_i^2\right)X^{2p-2}\right]\le\left\|\sum_{i=1}^{n}H_i^2\right\|\operatorname{Tr}X^{2p-2}=\sigma^2\operatorname{Tr}X^{2p-2}, \tag{46}$$

which gives

$$\mathbb{E}\operatorname{Tr}X^{2p}\le\sigma^2(2p-1)\mathbb{E}\operatorname{Tr}X^{2p-2}. \tag{47}$$

Applying this inequality, recursively, we get

$$\mathbb{E}\operatorname{Tr}X^{2p}\le\left[(2p-1)(2p-3)\cdots(3)(1)\right]\sigma^{2p}\mathbb{E}\operatorname{Tr}X^{0}=(2p-1)!!\sigma^{2p}d$$

Hence,

$$\mathbb{E}\|X\|\le\left(\mathbb{E}\operatorname{Tr}X^{2p}\right)^{\frac{1}{2p}}\le\left[(2p-1)!!\right]^{\frac{1}{2p}}\sigma d^{\frac{1}{2p}}.$$

Taking $p=\lceil\log d\rceil$ and using the fact that $(2p-1)!!\le\left(\frac{2p+1}{e}\right)^p$ (see [Tro15a] for an elementary proof consisting essentially of taking logarithms and comparing the sum with an integral) we get

$$\mathbb{E}\|X\|\le\left(\frac{2\lceil\log d\rceil+1}{e}\right)^{\frac{1}{2}}\sigma d^{\frac{1}{2\lceil\log d\rceil}}\le\left(2\lceil\log d\rceil+1\right)^{\frac{1}{2}}\sigma.$$

□

**Remark 4.31** *A similar argument can be used to prove Theorem 4.14 (the gaussian series case) based on gaussian integration by parts, see Section 7.2. in [Tro15c].*

**Remark 4.32** *Note that, up until the step from* (44) *to* (45) *all steps are equalities suggesting that this step may be the lossy step responsible by the suboptimal dimensional factor in several cases (although* (46) *can also potentially be lossy, it is not uncommon that* $\sum H_i^2$ *is a multiple of the identity matrix, which would render this step also an equality).*

*In fact, Joel Tropp [Tro15c] recently proved an improvement over the NCK inequality that, essentially, consists in replacing inequality* (45) *with a tighter argument. In a nutshell, the idea is that, if the* $H_i$*'s are non-commutative, most summands in* (44) *are actually expected to be smaller than the ones corresponding to* $q=0$ *and* $q=2p-2$*, which are the ones that appear in* (45).

### 4.7 Other Open Problems

#### 4.7.1 Oblivious Sparse Norm-Approximating Projections

There is an interesting random matrix problem related to Oblivious Sparse Norm-Approximating Projections [NN], a form of dimension reduction useful for fast linear algebra. In a nutshell, The idea is to try to find random matrices $\Pi$ that achieve dimension reduction, meaning $\Pi \in \mathbb{R}^{m\times n}$ with $m \ll n$, and that preserve the norm of every point in a certain subspace [NN], moreover, for the sake of computational efficiency, these matrices should be sparse (to allow for faster matrix-vector multiplication). In some sense, this is a generalization of the ideas of the Johnson-Lindenstrauss Lemma and Gordon's Escape through the Mesh Theorem that we will discuss next Section.

**Open Problem 4.4 (OSNAP [NN])** *Let $s \le d \le m \le n$.*

1. *Let $\Pi \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. entries*

$$\Pi_{ri} = \frac{\delta_{ri}\sigma_{ri}}{\sqrt{s}},$$

*where $\sigma_{ri}$ is a Rademacher random variable and*

$$\delta_{ri} = \begin{cases} \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{m} \\ 0 & \text{with probability} \quad 1-\frac{s}{m} \end{cases}$$

*Prove or disprove: there exist positive universal constants $c_1$ and $c_2$ such that*

*For any $U \in \mathbb{R}^{n\times d}$ for which $U^TU = I_{d\times d}$*

$$\operatorname{Prob}\left\{ \left\| (\Pi U)^T(\Pi U) - I \right\| \ge \varepsilon \right\} < \delta,$$

*for $m \ge c_1 \frac{d+\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \ge c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.*

2. *Same setting as in (1) but conditioning on*

$$\sum_{r=1}^m \delta_{ri} = s, \qquad \text{for all } i,$$

*meaning that each column of $\Pi$ has exactly $s$ non-zero elements, rather than on average. The conjecture is then slightly different:*

*Prove or disprove: there exist positive universal constants $c_1$ and $c_2$ such that*

*For any $U \in \mathbb{R}^{n\times d}$ for which $U^TU = I_{d\times d}$*

$$\operatorname{Prob}\left\{ \left\| (\Pi U)^T(\Pi U) - I \right\| \ge \varepsilon \right\} < \delta,$$

*for $m \ge c_1 \frac{d+\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \ge c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon}$.*

3. *The conjecture in (1) but for the specific choice of $U$:*

$$U = \begin{bmatrix} I_{d\times d} \\ 0_{(n-d)\times d} \end{bmatrix}.$$

*In this case, the object in question is a sum of rank 1 independent matrices. More precisely, $z_1, \ldots, z_m \in \mathbb{R}^d$ (corresponding to the first $d$ coordinates of each of the $m$ rows of $\Pi$) are i.i.d. random vectors with i.i.d. entries*

$$(z_k)_j = \begin{cases} -\frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \\ 0 & \text{with probability} \quad 1 - \frac{s}{m} \\ \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \end{cases}$$

*Note that $\mathbb{E} z_k z_k^T = \frac{1}{m} I_{d\times d}$. The conjecture is then that, there exists $c_1$ and $c_2$ positive universal constants such that*

$$\text{Prob}\left\{ \left\| \sum_{k=1}^m \left[ z_k z_k^T - \mathbb{E} z_k z_k^T \right] \right\| \geq \varepsilon \right\} < \delta,$$

*for $m \geq c_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.*

*I think this would is an interesting question even for fixed $\delta$, for say $\delta = 0.1$, or even simply understand the value of*

$$\mathbb{E} \left\| \sum_{k=1}^m \left[ z_k z_k^T - \mathbb{E} z_k z_k^T \right] \right\|.$$

### 4.7.2 $k$-lifts of graphs

Given a graph $G$, on $n$ nodes and with max-degree $\Delta$, and an integer $k \geq 2$ a random $k$ lift $G^{\otimes k}$ of $G$ is a graph on $kn$ nodes obtained by replacing each edge of $G$ by a random $k \times k$ bipartite matching. More precisely, the adjacency matrix $A^{\otimes k}$ of $G^{\otimes k}$ is a $nk \times nk$ matrix with $k \times k$ blocks given by

$$A_{ij}^{\otimes k} = A_{ij} \Pi_{ij},$$

where $\Pi_{ij}$ is uniformly randomly drawn from the set of permutations on $k$ elements, and all the edges are independent, except for the fact that $\Pi_{ij} = \Pi_{ji}$. In other words,

$$A^{\otimes k} = \sum_{i<j} A_{ij} \left( e_i e_j^T \otimes \Pi_{ij} + e_j e_i^T \otimes \Pi_{ij}^T \right),$$

where $\otimes$ corresponds to the Kronecker product. Note that

$$\mathbb{E} A^{\otimes k} = A \otimes \left( \frac{1}{k} J \right),$$

where $J = \mathbf{1}\mathbf{1}^T$ is the all-ones matrix.

**Open Problem 4.5 (Random $k$-lifts of graphs)** *Give a tight upperbound to*

$$\mathbb{E}\left\| A^{\otimes k} - \mathbb{E}A^{\otimes k} \right\|.$$

Oliveira [Oli10] gives a bound that is essentially of the form $\sqrt{\Delta \log(nk)}$, while the results in [ABG12] suggest that one may expect more concentration for large $k$. It is worth noting that the case of $k = 2$ can essentially be reduced to a problem where the entries of the random matrix are independent and the results in [BvH15] can be applied to, in some case, remove the logarithmic factor.

## 4.8 Another open problem

Feige [Fei05] posed the following remarkable conjecture (see also [Sam66, Sam69, Sam68])

**Conjecture 4.33** *Given $n$ independent random variables $X_1, \dots, X_n$ s.t., for all $i$, $X_i \geq 0$ and $\mathbb{E}X_i = 1$ we have*

$$\text{Prob}\left( \sum_{i=1}^{n} X_i \geq n + 1 \right) \leq 1 - e^{-1}$$

Note that, if $X_i$ are i.i.d. and $X_i = n + 1$ with probability $1/(n + 1)$ and $X_i = 0$ otherwise, then $\text{Prob}\left( \sum_{i=1}^{n} X_i \geq n + 1 \right) = 1 - \left( \frac{n}{n+1} \right)^n \approx 1 - e^{-1}$.

**Open Problem 4.6** *Prove or disprove Conjecture 4.33.*[21]

[21] We thank Francisco Unda and Philippe Rigollet for suggesting this problem.