Lecture #2: Friday, 7 Feb, 2025
Lecturer: Pravesh Kothari
Scribe: Nilava Metya

1 Khintchine's inequality

We aim to prove the matrix Khintchine's inequality. But before that, let's look at the same for scalar random variables.

Theorem 1
Let $a_1, \cdots, a_n \in \mathbb{R}$ and $t \in \mathbb{N}$. Then

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} g_i a_i\right)^{2t}\right]^{\frac{1}{2t}} \leq \sqrt{2t-1}\sqrt{\sum_{i=1}^{n} a_i^2}$$

where the expectation is taken over independent Gaussians $g_1, \cdots, g_n \sim \mathcal{N}(0,1)$.

Proof. Denote $\boldsymbol{a} = (a_1, \cdots, a_n), \boldsymbol{g} = (g_1, \cdots, g_n)$. Observe that $g \coloneqq \dfrac{\langle \boldsymbol{g}, \boldsymbol{a} \rangle}{\|\boldsymbol{a}\|_2} \sim \mathcal{N}(0,1)$ if g_i's are independent standard normals. Recall that $\mathbb{E}\left[g^{2t}\right] = \prod_{i=1}^{t}(2i-1) \leq (2t-1)^t$ for such g. The former equality has been proven in Appendix A.1. Multiplying throughout by $\|\boldsymbol{a}\|_2^{2t}$ and taking $(2t)^{\text{th}}$ square roots gives the desired result. ∎

We now move on to Khintchine's inequality for matrices whose proof relies on a bound stated in Lemma 10 in Appendix A.2.

Theorem 2
$\exists c > 0$ such that $\forall$ symmetric $A_1, \cdots, A_n \in \mathbb{R}^{m \times m}$ (with $m \geq 3$) we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} g_i A_i\right\|_2\right] \leq c\sqrt{\lg m}\left\|\sum_{i=1}^{n} A_i^2\right\|_2^{\frac{1}{2}}$$

where the expectation is taken over independent Gaussians $g_1, \cdots, g_n \sim \mathcal{N}(0,1)$.

Proof. Denote $X := \sum_{i=1}^{n} g_i A_i$. Then X is symmetric and $\|X\|_2$ is the absolute value of its largest sized eigenvalue. Say X has eigenvalues $\lambda_1, \cdots, \lambda_m \in \mathbb{R}$ such that $|\lambda_m| \geq |\lambda_i| \ \forall\ i \in [m]$. Let $\sigma = \left\|\sum_{i=1}^{n} A_i^2\right\|_2$ which is also the absolute value of the eigenvalue of $\sum A_i^2$ with largest size. Then for any $t \in \mathbb{N}$ we have $\|X\|_2^{2t} = |\lambda_m|^{2t} \leq \sum_{i=1}^{m} \lambda_i^{2t} = \mathrm{Tr}(X^{2t})$, whence

$$\begin{aligned}
\mathbb{E}\left[\|X\|_2\right] &\leq \mathbb{E}\left[\left(\mathrm{Tr}(X^{2t})\right)^{\frac{1}{2t}}\right] \\
&\overset{\text{Jensen}}{\leq} \left(\mathbb{E}\left[\mathrm{Tr}(X^{2t})\right]\right)^{\frac{1}{2t}} \\
&\overset{\text{Lemma 10}}{\leq} \sqrt{2t-1}\left[\mathrm{Tr}\left[\left(\sum_{i=1}^{n} A_i^2\right)^t\right]\right]^{\frac{1}{2t}} \\
&\leq \sqrt{2t-1}\left(m\sigma^{2t}\right)^{\frac{1}{2t}} = \sqrt{2t-1}\sigma m^{\frac{1}{2t}}.
\end{aligned}$$

Choosing $t \in \left[\frac{\lg m - 1}{2}, \frac{\lg m + 1}{2}\right] \cap \mathbb{Z}$ gives $\mathbb{E}\left[\|X\|_2\right] \leq \sigma\sqrt{\lg m} \cdot \underbrace{2^{\frac{1}{2}+\frac{1}{\lg(9/4)}}}_{c}$. ■

Now let's move on to Khintchine's inequality with Rademacher weights, instead of Gaussian. We expect such a statement to be true because somehow the 'tails' in the Rademacher case are bounded above by Gaussians. What we precisely want is stated in Appendix A.3.

Theorem 3
$\exists c > 0$ such that $\forall$ symmetric $A_1, \cdots, A_n \in \mathbb{R}^{m \times m}$, we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} b_i A_i\right\|_2\right] \leq c\sqrt{\lg m}\left\|\sum_{i=1}^{n} A_i^2\right\|_2^{\frac{1}{2}}$$

where the expectation is taken over a uniform draw of $\boldsymbol{b} \in \{\pm 1\}^n$.

Proof. Immediate from Theorem 2 and Lemma 11 (different c from Theorem 2). ■

2 Sums of independent random variables

We now want to look at quantities like $\left\|\sum_{i=1}^{n}(H_i - \mathbb{E}\left[H_i\right])\right\|_2$ where the H_i's are chosen uniformly from the set of $m \times m$ real symmetric matrices. In such cases, a symmetrization trick is really useful which is presented in Appendix A.4. We shall use this to prove the following expected deviation.

Theorem 4
$\exists\ c > 0$ such that if $H_1, \cdots, H_n \in \mathbb{R}^{m\times m}$ are symmetric and chosen uniformly randomly, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\left\|\sum_{i=1}^{n}(H_i - \mathbb{E}[H_i])\right\|_2\right] \le c\sqrt{\lg m}\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\left\|\sum_i H_i^2\right\|_2^{\frac{1}{2}}\right].$$

Proof. Lemma 12 says that the LHS is at most $2\mathop{\mathbb{E}}_{\substack{\boldsymbol{b}\in\{\pm1\}^n\\ \boldsymbol{H}}}\left\|\sum_{i=1}^{n} b_i H_i\right\|_2$. But this is the same as $2\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{b}\in\{\pm1\}^n}\left(\left\|\sum_{i=1}^{n} b_i H_i\right\|_2 \;\middle|\; H_1, \cdots, H_n\right)\right]$ by the law of total expectation. By Theorem 3, this expression is at most $\underbrace{2c'}_{c}\sqrt{\lg m}\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{b}\in\{\pm1\}^n}\left(\left\|\sum_{i=1}^{n} H_i^2\right\|_2^{\frac{1}{2}} \;\middle|\; H_1, \cdots, H_n\right)\right] = c\sqrt{\lg m}\mathop{\mathbb{E}}_{\boldsymbol{H}}\left\|\sum_{i=1}^{n} H_i^2\right\|_2^{\frac{1}{2}}$. ■

2.1 Chernoff bound for random symmetric matrices

Theorem 5
$\exists c, c' > 0$ such that if $H_1, \cdots, H_n \in \mathbb{R}^{m\times m}$ are symmetric and chosen uniform and independently such that $H_i \succeq 0$ a.s. and $H_i \preceq M\mathbb{I}$ a.s., we have the following $\forall\ \varepsilon > 0$:

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{n} H_i\right\|_2\right] &\le (1+\varepsilon)\left\|\sum \mathbb{E}[H_i]\right\|_2 + \left(1+\frac{1}{\varepsilon}\right)cM\sqrt{\lg m}\\ &\overset{\text{if } \lambda_{\max}(\sum\mathbb{E}[H_i])\gg M\lg m}{\le} (1+\mathcal{O}(\varepsilon))\left\|\sum_{i=1}^{n}\mathbb{E}[H_i]\right\|_2.\end{aligned}$$

Proof. First write $\mathbb{E}\left[\left\|\sum_{i=1}^{n} H_i\right\|_2\right] \le \mathbb{E}\left[\left\|\sum_i H_i - \mathbb{E}[H_i]\right\|_2\right] + \left\|\sum_i \mathbb{E}[H_i]\right\|_2$ by the triangle inequality. Then $\mathbb{E}\left[\left\|\sum_i H_i - \mathbb{E}[H_i]\right\|_2\right] \overset{\text{Theorem 4}}{\lesssim} \sqrt{\log m}\mathbb{E}\left[\left\|\sum_i H_i^2\right\|_2^{\frac{1}{2}}\right] \overset{0\preceq H_i\preceq MI}{\lesssim} \sqrt{\log m}\cdot\mathbb{E}\left[\sqrt{M}\left\|\sum_i H_i\right\|_2^{\frac{1}{2}}\right] \le \sqrt{M\log m}\mathbb{E}\left[\left\|\sum_i H_i\right\|_2\right]^{\frac{1}{2}}$. Taking $x := \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n} H_i\right\|_2\right]}$

gives from the last inequality that $x^2 - xc\sqrt{M\log m} - \|\sum \mathbb{E}[H_i]\|_2 \leq 0$. Since this is a (strictly) convex quadratic, we must have $x \leq \dfrac{c\sqrt{M\log m} + \sqrt{c^2 M\log m + 4\|\sum \mathbb{E}[H_i]\|_2}}{2}$ for the previous inequality to be true. This in turn implies that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum H_i\right\|_2\right] = x^2 &\leq \frac{1}{2}c^2 M\log m + \left\|\sum \mathbb{E}[H_i]\right\|_2 + \\
&\qquad \frac{1}{2}\sqrt{\frac{1}{\varepsilon}c^2 M\log m}\cdot\sqrt{\varepsilon}\sqrt{c^2 M\log m + 4\left\|\sum \mathbb{E}[H_i]\right\|} \\
&\overset{\text{GM}\leq\text{AM}}{\leq} \frac{1}{2}c^2 M\log m + \left\|\sum \mathbb{E}[H_i]\right\|_2 + \\
&\qquad \frac{\frac{1}{\varepsilon}c^2 M\log m + \varepsilon c^2 M\log m + 4\varepsilon\|\sum \mathbb{E}[H_i]\|_2}{4} \\
&= \frac{1}{4}\left(2 + \varepsilon + \frac{1}{\varepsilon}\right)c^2 M\log m + (1+\varepsilon)\left\|\sum \mathbb{E}[H_i]\right\|_2.
\end{aligned}$$

The above is true for any $\varepsilon > 0$. Doing the same arugment with ε replaced by $\frac{1}{\varepsilon}$ gives a coefficient of $\left(1 + \frac{1}{\varepsilon}\right)$ in the second term (first term is unchanged), hence we can assume $\varepsilon < 1$. Thus, the above inequality for $\varepsilon \in (0,1)$ can be stated as

$$\mathbb{E}\left[\left\|\sum H_i\right\|_2\right] \leq \left(1 + \frac{1}{\varepsilon}\right)c' M\log m + (1+\varepsilon)\left\|\sum \mathbb{E}[H_i]\right\|_2.$$

■

2.2 Bernstein bound for random symmetric matrices

Theorem 6
$\exists c_1, c_2 > 0$ such that if $H_1, \cdots, H_n \in \mathbb{R}^{m\times m}$ are chosen uniformly randomly from the space of real symmetric matrices with $\mathbb{E}[H_i] = 0$ and $\|H_i\|_2 \leq M$ almost surely $\forall\ i$, then

$$\mathbb{E}\left[\left\|\sum H_i\right\|_2\right] \leq c_1\sqrt{\log m}\left\|\sum \mathbb{E}\left[H_i^2\right]\right\|_2^{\frac{1}{2}} + c_2 M\log m.$$

Proof. $\mathbb{E}\left[\left\|\sum H_i\right\|_2\right] \overset{\text{Theorem 4}}{\leq} c\sqrt{\log m}\,\mathbb{E}\left[\left\|\sum H_i^2\right\|_2^{\frac{1}{2}}\right] \overset{\text{Jensen}}{\leq} c\sqrt{\log m}\,\mathbb{E}\left[\left\|\sum H_i^2\right\|_2\right]^{\frac{1}{2}} \overset{\text{Theorem 5}}{\leq}$ $c\sqrt{\log m}\left(c'\left\|\sum \mathbb{E}\left[H_i^2\right]\right\|_2^{\frac{1}{2}} + c'' M\sqrt{\log m}\right)$ where the last inequality is true by replacing H_i, M with H_i^2, M^2 in Theorem 5 and then applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ and end by taking $c_1 := cc', c_2 := cc''$. ■

3 Applications

3.1 Another proof of Graph Sparsification

Problem statement and algorithm. Let's recall the problem of graph sparsification. Given an unweighted undirected graph $G = (V, E)$ with $n = |V|$ we want to find another undirected *weighted* graph $G'(V', E', w)$ with $E' \subseteq E$ such that (with multiplicative error) $\sum_{e \in E'} w_e L_e \overset{1\pm\varepsilon}{\simeq} \sum_{e \in E} L_e$ (so that cut-values are preserved), where the edge Laplacians are $L_{\{i,j\}} := \delta_{ii} + \delta_{jj} - \delta_{ij} - \delta_{ji}$ where $\delta_{\alpha\beta}$ is the $n \times n$ matrix with a 1 entry in position (α, β) and zero elsewhere. Recall the Laplacian of the graph $(L =)L_G = \sum_{e \in E} L_e$ and the normalized edge Laplacian $\tilde{L}_e = L^{\dagger/2} L_e L^{\dagger/2}$. With this new notation, our goal is the same as $\sum_{e \in E'} w_e \tilde{L}_e \overset{1\pm\varepsilon}{\simeq} \mathbb{I}_\perp$. Look at Lecture 1 for details.

The ideal algorithm was based on importance sampling, to independently sample edge $e \in E$ with probability $p_e := \frac{1}{R} \left\| \tilde{L}_e \right\|_2$ to be included in E', where R is to be chosen later. If $e \in E'$ we will set $w_e = \frac{1}{p_e}$. Then $\mathbb{E}\left[|E'|\right] = \sum_{e \in E} p_e = \sum_{e \in E} \frac{1}{R} \left\| \tilde{L}_e \right\|_2 \overset{(*)}{=} \sum_{e \in E} \frac{1}{R} \operatorname{Tr}\left(\tilde{L}_e\right) = \frac{1}{R} \operatorname{Tr}\left(\boldsymbol{I}_\perp\right) = \frac{n-1}{R}$. $(*)$ follows from the fact that each $\tilde{L}_e \succeq 0$ and has rank 1. To prove that $|E'|$ is concentrated around this number, we can use the usual Chernoff bound.

Error analysis. Let $Q_e := \begin{cases} \frac{1}{p_e} \tilde{L}_e & \text{if } e \in E' \\ \mathbf{0}_{n \times n} & \text{otherwise} \end{cases} = \underbrace{\mathbf{1}[e \text{ sampled}] \cdot \frac{1}{p_e}}_{w_e} \tilde{L}_e$ for each $e \in E$ (if an edge is not present in E', then think of its weight as 0). Each Q_e is positive semidefinite and has all eigenvalues $\le R$. So $\mathbb{E}\left[Q_e\right] = \tilde{L}_e$. So $\mathbb{E}\left[\left\| \sum_{e \in E'} w_e \tilde{L}_e \right\|_2\right] = \mathbb{E}\left[\left\| \sum_{e \in E} Q_e \right\|_2\right] \le (1+\varepsilon) \left\| \sum_{e \in E} \mathbb{E}\left[Q_e\right] \right\|_2 + \left(1 + \frac{1}{\varepsilon}\right) cR \log n$ where the last inequality follows from Theorem 5. Choosing $R = \frac{\varepsilon^2}{\log n}$ and noting that $\left\| \sum_{e \in E} \mathbb{E}\left[Q_e\right] \right\|_2 = \left\| \sum_{e \in E} \tilde{L}_e \right\|_2 = \|\boldsymbol{I}_\perp\|_2 = 1$ guarantees $\mathbb{E}\left[\left\| \sum_{e \in E'} w_e \tilde{L}_e \right\|_2\right] \le 1 + \mathcal{O}(\varepsilon)$.

3.2 'Machine Learning'

We have an (unknown) distribution D on R^d with a random variable $X \sim D$ satisfying $\mathbb{E}_D\left[X\right] = 0, \mathbb{E}_D\left[XX^\top\right] = \Sigma$ (Σ also unknown). To *learn* Σ, an intuitive algorithm is to sample $X_1, \cdots, X_m \sim D$ independently and report an estimate $\hat{\Sigma} \coloneqq \frac{1}{m}\sum_{i=1}^{m} X_i X_i^\top$. We want to get theoretical guarantees of the form "$\left\|\hat{\Sigma} - \Sigma\right\|_2 \leq \varepsilon$ with probability 99%". For 'real life' cases, we assume Σ is invertible. Since the above estimate can depend on the eigenvalues of Σ, we instead demand a relative estimate so that people can have an overall idea of ε's are good enough. So we instead demand a guarantee of the form "$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \boldsymbol{I}\right\|_2 \leq \varepsilon$ with probability 99%". This is equivalent to demanding "$\boldsymbol{v}^\top \left(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\right) \boldsymbol{v} \in (1-\varepsilon, 1+\varepsilon) \left\|\boldsymbol{v}\right\|^2 \ \forall\, \boldsymbol{v} \in \mathbb{R}^d$ with probability 99%".

Comparing Theorem 6 with the above expression forces us to write it as $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} Y_i Y_i^\top$ where $Y_i \coloneqq \Sigma^{-1/2} X_i$. Noting that $\mathbb{E}\left[Y_i Y_i^\top\right] = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \boldsymbol{I}$ suggests that we can use Theorem 6 with $H_i \coloneqq \frac{1}{m}\left(Y_i Y_i^\top - \boldsymbol{I}\right)$, provided we know some K such that $K \geq \left\|\Sigma^{-1/2} X\right\|_2$ almost surely. The latter assumption guarantees $\|H_i\| \leq \frac{1}{m}\left(\|Y_i\|^2 + 1\right) \lesssim \frac{K^2}{m}$. Using the raw form of Bernouli (Theorem 4) gives $\mathbb{E}\left[\left\|\sum H_i\right\|_2\right] \leq c\sqrt{\log d} \cdot \dfrac{K^2}{m}\sqrt{m} = cK^2\sqrt{\dfrac{\log d}{m}}$. Choosing $m > \frac{c^2 K^4 \log d}{\varepsilon^2}$ samples does the job.

A Appendix

A.1 Even moments of Gaussians

Proposition 7 (Gaussian integration by parts)
If $X \sim \mathcal{N}(0,1)$ and f is an differentiable function such that its derivative is integrable then $\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)]$.

Lemma 8 (even Gaussian moments)

If $X \sim \mathcal{N}(0,1)$ and $t \in \mathbb{N}$ then $\mathbb{E}\left[X^{2t}\right] = \prod_{i=1}^{t}(2i-1)$.

Proof. Taking $f(x) = x^{2t-1}$ in Proposition 7 gives $\mathbb{E}\left[X^{2t}\right] = \mathbb{E}[X \cdot f(X)] = (2t-1)\mathbb{E}\left[X^{2t-2}\right]$. This is one step in our calculation. Doing this recursively gives $\mathbb{E}\left[X^{2t}\right] = \prod_{i=2}^{t}(2i-1) \cdot \mathbb{E}\left[X^2\right] = \prod_{i=1}^{t}(2i-1)$ because $\mathbb{E}[X^2] = \mathrm{Var}[X] = 1$. ■

A.2 Bounding expectation of trace of even powers of Gaussian linear combination of matrices

Lemma 9
Let $X, A_1, \cdots, A_n \in \mathbb{R}^{m\times m}$ be symmetric matrices, $s \in \mathbb{N}$. Then $\mathrm{Tr}(A_k X^l A_k X^{2s-l}) \leq \mathrm{Tr}(A_k^2 X^{2s})$ for every $k \in [n], l \in [2s] \cup \{0\}$.

Proof. Consider the function $f(k,l) = \mathrm{Tr}(A_i X^l A_i X^{2s-l})$ for each $k \in [n], l \in [2s] \cup \{0\}$. X is symmetric, hence it can be decomposed as $X = \sum_{i=1}^{m} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ where $\{\boldsymbol{v}_i\}_{i=1}^m$ is an orthonormal basis of $\mathbb{R}^m$, and $\lambda_i \in \mathbb{R}$ are the eigenvalues of X. Then

$$\begin{aligned}
f(k,l) &= \mathrm{Tr}(A_k X^l A_k X^{2s-l}) \\
&= \mathrm{Tr}\left(A_k \left(\sum_{i=1}^{m} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top\right)^l A_k \left(\sum_{j=1}^{m} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top\right)^{2s-l}\right) \\
&= \mathrm{Tr}\left(A_k \left(\sum_{i=1}^{m} \lambda_i^l \boldsymbol{v}_i \boldsymbol{v}_i^\top\right) A_k \left(\sum_{j=1}^{m} \lambda_j^{2s-l} \boldsymbol{v}_j \boldsymbol{v}_j^\top\right)\right) \\
&= \sum_{i,j} \lambda_i^l \lambda_j^{2s-l}\, \mathrm{Tr}\left(A_k \boldsymbol{v}_i \boldsymbol{v}_i^\top A_k \boldsymbol{v}_j \boldsymbol{v}_j^\top\right) \\
&\leq \sum_{i,j} |\lambda_i|^l \left|\lambda_j\right|^{2s-l} \mathrm{Tr}\left(A_k \boldsymbol{v}_i \boldsymbol{v}_i^\top A_k \boldsymbol{v}_j \boldsymbol{v}_j^\top\right)
\end{aligned}$$

The expression $\sum_{i,j} |\lambda_i|^l |\lambda_j|^{2s-l} \operatorname{Tr}\left(A_k \boldsymbol{v}_i \boldsymbol{v}_i^\top A_k \boldsymbol{v}_j \boldsymbol{v}_j^\top\right)$ is convex in l. Hence it maximizes at the endpoints of the range of l, namely $0, 2s$. Due to the symmetry of A_k, the value of the expression at $l = 0$ is the same as that at $l = 2s$. This common value is $\sum_{i,j} \lambda_j^{2s} \operatorname{Tr}\left(A_k \boldsymbol{v}_i \boldsymbol{v}_i^\top A_k \boldsymbol{v}_j \boldsymbol{v}_j^\top\right)$ which is the same as $f(k,0)$ (this step would not be possible if $2s$ were replaced by some odd number). Thus we have proven that for any $k \in [n], l \in [2s] \cup \{0\}$, $\operatorname{Tr}\left(A_k X^l A_k X^{2s-l}\right) = f(k,l) \le f(k,0) = \operatorname{Tr}(A_k^2 X^{2s})$. ■

Lemma 10
Let $A_1, \cdots, A_n \in \mathbb{R}^{n\times n}$ be symmetric matrices and $g_1, \cdots, g_n \sim \mathcal{N}(0,1)$ be independent Gaussians. Denote $X := \sum_{i=1}^n g_i A_i$. Then

$$\left(\mathbb{E}\left[\operatorname{Tr}\left(X^{2t}\right)\right]\right)^{\frac{1}{2t}} \le \sqrt{2t-1}\left[\operatorname{Tr}\left[\left(\sum_{i=1}^n A_i^2\right)^t\right]\right]^{\frac{1}{2t}}.$$

Proof. We will first use Proposition 7 similar to the proof of Lemma 8, and then invoke Lemma 9 with $s = t-1$. We have

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Tr}\left(X^{2t}\right)\right] &= \mathbb{E}\left[\operatorname{Tr}\left(X \cdot X^{2t-1}\right)\right] \\
&= \sum_{k=1}^n \mathbb{E}\left[g_k \operatorname{Tr}\left(A_k X^{2t-1}\right)\right] \\
&\overset{\text{Proposition 7}}{=} \sum_{k=1}^n \sum_{l=0}^{2t-2} \mathbb{E}\left[\operatorname{Tr}\left(A_k X^l \underbrace{\left(\partial_{g_k} X\right)}_{A_k} X^{2t-2-l}\right)\right] \\
&= \sum_{k=1}^n \sum_{l=0}^{2t-2} \mathbb{E}\left[\operatorname{Tr}\left(A_k X^l A_k X^{2t-2-l}\right)\right] \\
&\overset{\text{Lemma 9}}{\le} \sum_{l=0}^{2t-2} \sum_{k=1}^n \mathbb{E}\left[\operatorname{Tr}\left(A_k^2 X^{2t-2}\right)\right] \\
&= \sum_{l=0}^{2t-2} \mathbb{E}\left[\operatorname{Tr}\left(\left(\sum_{k=1}^n A_k^2\right) X^{2t-2}\right)\right] \\
&= (2t-1)\mathbb{E}\left[\operatorname{Tr}\left(\left(\sum_{k=1}^n A_k^2\right) X^{2t-2}\right)\right] \\
&\overset{\text{Hölder}}{\le} (2t-1)\mathbb{E}\left[\left[\operatorname{Tr}\left(\left(\sum_k A_k^2\right)^t\right)\right]^{\frac{1}{t}} \cdot \left(\operatorname{Tr}(X^{2t})\right)^{1-\frac{1}{t}}\right]
\end{aligned}$$

$$= (2t-1)\left[\mathrm{Tr}\left(\left(\sum_k A_k^2\right)^t\right)\right]^{\frac{1}{t}} \mathbb{E}\left[\left(\mathrm{Tr}(X^{2t})\right)^{1-\frac{1}{t}}\right]$$

$$\overset{\text{Jensen}}{\leq} (2t-1)\left[\mathrm{Tr}\left(\left(\sum_k A_k^2\right)^t\right)\right]^{\frac{1}{t}} \left(\mathbb{E}\left[\mathrm{Tr}(X^{2t})\right]\right)^{1-\frac{1}{t}}$$

$$\implies \left(\mathbb{E}\left[\mathrm{Tr}(X^{2t})\right]\right)^{\frac{1}{t}} \leq (2t-1)\left[\mathrm{Tr}\left(\left(\sum_k A_k^2\right)^t\right)\right]^{\frac{1}{t}}$$

$$\implies \left(\mathbb{E}\left[\mathrm{Tr}(X^{2t})\right]\right)^{\frac{1}{2t}} \leq \sqrt{2t-1}\left[\mathrm{Tr}\left(\left(\sum_k A_k^2\right)^t\right)\right]^{\frac{1}{2t}}.$$

■

A.3 Gauss dominates Rademacher

Lemma 11
If $\boldsymbol{b}$ is uniformly random on $\{\pm 1\}^n$ and $g_1, \cdots, g_n \sim \mathcal{N}(0,1)$ are independent (and also independent of $\boldsymbol{b}$), then

$$\mathop{\mathbb{E}}_{\boldsymbol{b}} \left\|\sum_i b_i A_i\right\|_2 \leq \sqrt{\frac{\pi}{2}} \mathop{\mathbb{E}}_{\boldsymbol{g}} \left\|\sum_i g_i A_i\right\|_2.$$

Proof. $\mathop{\mathbb{E}}_{\boldsymbol{g}} \left\|\sum_i g_i A_i\right\|_2 = \mathop{\mathbb{E}}_{\boldsymbol{g},\boldsymbol{b}} \left\|\sum_i |g_i|\, b_i A_i\right\|_2 \geq \mathop{\mathbb{E}}_{\boldsymbol{b}} \left\|\sum_i b_i \mathop{\mathbb{E}}_{\boldsymbol{g}} |g_i|\, A_i\right\|_2 = \sqrt{\frac{2}{\pi}} \mathop{\mathbb{E}}_{\boldsymbol{b}} \left\|\sum_i b_i A_i\right\|_2.$ ■

A.4 Symmetrization trick with Rademacher

Lemma 12
If $H_1, \cdots, H_n \in \mathbb{R}^{m\times m}$ are chosen uniformly randomly and independently from the set of $m \times m$ real symmetric matrices, then

$$\mathop{\mathbb{E}}_{\boldsymbol{H}} \left[\left\|\sum_{i=1}^n (H_i - \mathbb{E}[H_i])\right\|_2\right] \leq 2 \mathop{\mathbb{E}}_{\substack{\boldsymbol{b}\in\{\pm1\}^n \\ \boldsymbol{H}}} \left[\left\|\sum_{i=1}^n b_i H_i\right\|_2\right].$$

Proof. The symmetrization trick is to introduce an independent identical copy of each H_i, namely H_i', then each $H_i - \mathbb{E}[H_i]$ becomes $\mathop{\mathbb{E}}_{\boldsymbol{H}'}[H_i - H_i']$ because $\mathbb{E}[H_i] = \mathbb{E}[H_i']$. Then starting from the LHS we have,

$$\mathop{\mathbb{E}}_{\boldsymbol{H}} \left[\left\|\sum_{i=1}^n (H_i - \mathbb{E}[H_i])\right\|_2\right] = \mathop{\mathbb{E}}_{\boldsymbol{H}} \left[\left\|\mathop{\mathbb{E}}_{\boldsymbol{H}'} \left[\sum_{i=1}^n H_i - H_i'\right\|_2\right]\right]$$

$$
\begin{aligned}
&\overset{\text{Jensen}}{\leq} \mathop{\mathbb{E}}_{\boldsymbol{H}} \mathop{\mathbb{E}}_{\boldsymbol{H}'} \left[\left\| \sum_i H_i - H_i' \right\|_2 \right] \\
&\overset{(*)}{=} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{H}'} \left[\mathop{\mathbb{E}}_{\boldsymbol{b}\in\{\pm 1\}^n} \left\| \sum_{i=1}^n b_i (H_i - H_i') \right\|_2 \right] \\
&\overset{(**)}{\leq} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{H}'} \left[\mathop{\mathbb{E}}_{\boldsymbol{b}\in\{\pm 1\}^n} \left\| \sum_{i=1}^n b_i H_i \right\|_2 \right] + \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{H}'} \left[\mathop{\mathbb{E}}_{\boldsymbol{b}\in\{\pm 1\}^n} \left\| \sum_{i=1}^n b_i H_i' \right\|_2 \right] \\
&= 2 \mathop{\mathbb{E}}_{\substack{\boldsymbol{b}\in\{\pm 1\}^n \\ \boldsymbol{H}}} \left\| \sum_{i=1}^n b_i H_i \right\|_2 .
\end{aligned}
$$

$\overset{(*)}{=}$ holds because $H_i - H_i'$ has a symmetric distribution and hence $b_i(H_i - H_i') \overset{D}{\equiv} H_i - H_i'$.
$\overset{(**)}{\leq}$ is true by triangle inequality on $\|\cdot\|_2$. ■