# Lecture notes for "PIMS Summer School on Randomized Techniques for Combinatorial Algorithms"

Nicholas J. A. Harvey*

August 19, 2014

**Abstract**

In this lecture we will discuss two distinct topics: graph sparsification, and concentration bounds for sums of random matrices.

## 1 Graph Sparsification

The first result we discuss is graph sparsification: approximating a graph by weighted subgraphs of itself. These techniques have been used to design fast algorithms for combinatorial or linear algebraic problems, as a rounding technique in approximation algorithms, and to derive strong results in pure mathematics.

### 1.1 Notation

Let $[k] = \{1, ..., k\}$. The value $L$ will denote a universal constant whose value is not the same at each occurrence.

Let $G = (V, E)$ be a graph and $u : E \to \mathbb{R}$. We will typically assume that $V = [n]$. Then $\mathrm{supp}(u)$ denotes $\{ e \in E : u_e \neq 0 \}$ and for any $F \subseteq E$, $u(F)$ denotes $\sum_{e \in F} u_e$. Also, for any $U \subseteq V$,

$$\delta(U) \;=\; \{ \text{ edge } st : \text{exactly one of } s \text{ and } t \text{ is in } U \}.$$

The minimum cut in the graph is

$$\min \{ u(\delta(U)) : \emptyset \neq U \subsetneq V \}. \tag{1.1}$$

*Portions of these notes are based on scribe notes written by Zachary Drudi.

## 2 The Random Sparsification Algorithm

Let $G = (V, E)$ be a graph with edge weights $u : E \to \mathbb{R}_{>0}$. We will let $n = |V|$. The goal of sparsification is to find a new vector of edge weights $w : E \to \mathbb{R}_{\geq 0}$ such that $\text{supp}(w)$ is small but yet $G$ with edge weights $u$ is "structurally similar" to $G$ with edge weights $w$. For our purposes, the "structural similarity" that we wish to guarantee is the following "cut preservation" condition:

$$(1-\epsilon) \cdot u(\delta(U)) \;\leq\; w(\delta(U)) \;\leq\; (1+\epsilon) \cdot u(\delta(U)) \qquad \forall U \subseteq V. \tag{2.1}$$

The random sparsification algorithm is shown in Algorithm 2.1. By using different sampling probabilities, one can prove different guarantees about the quality of the resulting sparsifier.

**Algorithm 2.1** The random graph sparsification algorithm.

**procedure** Sparsify($G$, $p$)

**input:** A graph $G = (V, E)$ with edge weights $u : E \to \mathbb{R}_{>0}$, sampling probabilities $p : E \to (0, 1]$ and a parameter $\rho$ that specifies the number of rounds of sampling.

**output:** Edge weights $w : E \to \mathbb{R}_{\geq 0}$.

Initially $w = 0$.

Let $(Z_{i,e})_{i \in [\rho],\, e \in E}$ be mutually independent, random variables in $\{0, 1\}$ with $\mathrm{E}[\, Z_{i,e} \,] = p_e$.

For $i = 1, ..., \rho$

    For each $e \in E$

        Increase $w_e$ by $Z_{i,e} \cdot (u_e / \rho p_e)$.

Return $w$

Note that $\mathrm{E}[\, w_e \,] = u_e$ for all $e \in E$ because, by linearity of expectation,

$$\mathrm{E}[\, w_e \,] \;=\; \sum_{i=1}^{\rho} \mathrm{E}[\, Z_{i,e} \,] \cdot (u_e / \rho p_e) \;=\; \sum_{i=1}^{\rho} u_e / \rho \;=\; u_e.$$

Furthermore, $\mathrm{E}[\, w(\delta(U)) \,] = u(\delta(U))$ for all $U \subseteq V$, again by linearity of expectation. So this algorithm preserves the weight of every edge and of every cut in expectation, regardless of the $p_e$ values.

## 3 Graph Skeletons

For simplicity let assume that $G$ is connected; if not, one can apply this argument to every connected component of $G$.

We will show that, as long as $\rho$ is logarithmic in $n$ and the sampling probabilities are at least the reciprocal of the minimum cut, then the cut preservation condition (2.1) is satisfied. The formal theorem is as follows.

**Theorem 3.1** (Karger [2, 3]). Let $K$ be the minimum cut value, as defined in (1.1). Suppose that

$$\begin{aligned} p_e &\geq u_e/K \qquad \forall e \in E \\ \rho &= L \log(n)/\epsilon^2 \end{aligned} \tag{3.1}$$

Then the weight vector $w$ output by Algorithm 2.1 satisfies (2.1) with high probability.

**Proof.** Consider a set of vertices $U \subseteq V$. Let $\mathcal{E}_U$ be the "bad event" that

$$w(\delta(U)) \notin [1-\epsilon, 1+\epsilon] \cdot u(\delta(U)).$$

Our aim is to show that, with high probability, $\mathcal{E}_U$ does not occur.

Without loss of generality we may assume $K = 1$. The reason is that multiplying each $u_e$ by a factor $\alpha$ also multiplies both $K$ and $w$ by $\alpha$, but does not affect our hypotheses on $p_e$ and $\rho$. So we may choose $\alpha = 1/K$ without affecting the hypotheses or conclusion of the theorem.

From the pseudocode in Algorithm 2.1, we have

$$w(\delta(U)) = \sum_{i \in [\rho]} \sum_{e \in \delta(U)} Z_{i,e} \cdot (u_e/\rho p_e).$$

This is a sum of independent random variables, each of which takes values in the bounded interval $[0, R]$, where $R = \max_e \frac{u_e}{\rho p_e} \leq 1/\rho$, due to (3.1) and the assumption that $K = 1$. We observed above that $\mu = \mathrm{E}[\, w(\delta(U))\,] = u(\delta(U))$. By a Chernoff bound (Theorem 5.13), the failure probability is

$$\begin{aligned} \Pr[\, \mathcal{E}_U\,] &\leq 2\exp\big(-\epsilon^2\mu/3R\big) \leq 2\exp\Big(-(\epsilon^2\rho/3)\cdot u(\delta(U))\Big) \\ &\leq \exp\Big(-L\log(n)\cdot u(\delta(U))\Big) = n^{-L\cdot u(\delta(U))}. \end{aligned} \tag{3.2}$$

This probability is at most $n^{-L}$ because the minimum cut value is $K = 1$, so $u(\delta(U)) \geq 1$. However, that analysis is too weak: ultimately we wish to bound

$$\Pr\left[\bigcup_{U \subseteq V} \mathcal{E}_U\right] \leq \sum_{U \subseteq V} \Pr[\, \mathcal{E}_U\,], \tag{3.3}$$

and bounding each term by $n^{-L}$ would not give a useful result as there are exponentially many terms, so the sum would exceed 1.

Fortunately most cuts admit a tighter bound on their failure probability. The following theorem shows that, in a quantitative sense, most cut values are much larger than $K = 1$.

**Theorem 3.2** (Karger). For any undirected graph with positive edge weights and min cut value 1,

$$|\{\, U \subseteq V \,:\, u(\delta(U)) \leq x\,\}| \leq n^{2x} \qquad \text{for all real } x \geq 1.$$

Thus, we may expand the right-hand side of (3.3) as

$$\begin{aligned}
\sum_{U \subseteq V} \Pr[\,\mathcal{E}_U\,] &= \sum_{\emptyset \neq U \subsetneq V} n^{-L \cdot u(\delta(U))} \qquad (3.4)\\
&= \sum_{s \geq 1} \sum_{\substack{U \subseteq V \\ s \leq u(\delta(U)) < s+1}} n^{-L \cdot u(\delta(U))} \\
&\leq \sum_{s \geq 1} n^{2 \cdot (s+1)} \cdot n^{-Ls} \qquad \text{(by Theorem 3.2 and (3.2))} \\
&\leq \sum_{s \geq 1} n^{-Ls} \;\leq\; n^{-L}.
\end{aligned}$$

In summary, we have shown that $\Pr\left[\bigcup_{U \subseteq V} \mathcal{E}_U\right] \leq n^{-L}$, completing the proof. ■

**Remarks**

- If the graph is unweighted (i.e., $u_e = 1$) and if $p_e = 1/K$, then random sparsification decreases the number of edges by a factor $\rho/K$ in expectation. This is because

$$\mathrm{E}[\,|\mathrm{supp}(w)|\,] \;=\; \sum_{e \in E} \Pr[\,w_e > 0\,] \;\leq\; \sum_{e \in E} \sum_{i \in [\rho]} \Pr[\,Z_{i,e} \neq 0\,] \;=\; |E| \cdot (\rho/K).$$

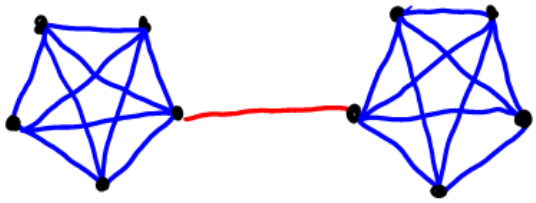
- Suppose we set each $p_e$ to be exactly $u_e/K$. Then the sparsifier is an "integral graph", in the sense that every edge weight $w_e$ is an integer multiple of $K/\rho$.
- Algorithms are known to compute $K$ in nearly-linear time.

**Applications**

- **Minimum $s$-$t$ cuts**. Given any algorithm to compute a minimum $s$-$t$ cut, instead of running it on the original unweighted graph $G$, we can run it on the graph with weights $w$. Return the resulting cut as an approximate min $s$-$t$ cut in $G$. This approach is faster by a factor of roughly $\rho/K$.
- **Sparsest cut, Max cut**. Similar ideas lead to a speedup of known approximation algorithms for these problems.

# 4 Graph sparsification by non-uniform sampling

Theorem 3.1 will produce a sparsified graph with few edges if $K$ is large, but it is less effective if $K$ is small. An example that highlights this issue is the "dumbbell graph", which consists of two disjoint cliques, each on $n/2$ vertices, and a single edge in the middle connecting the cliques. Here the minimum cut is $K = 1$, so Theorem 3.1 cannot achieve any sparsification of this graph.



Nevertheless, a natural idea is that we could run the sparsification algorithm separately on the two cliques (each of which has a large minimum cut value), and keep the edge in the middle as-is. Another way of saying this is that we should sample the middle edge with probability 1, but sample the clique edges with very low probability.

Generalizing this idea to arbitrary graphs, we would like to find some notion of how "important" each edge is. Should we sample the edge with low probability or high probability? Benczúr and Karger defined a notion of "edge strength" that gives a useful notion of "importance" for our purposes.

**Definition 4.1.** Let $G = (V, E)$ be a graph with edge weights $u : E \to \mathbb{R}_{>0}$. Let $K(H)$ denote the minimum cut value of the graph $H$, and let $G[T]$ denoted the subgraph of $G$ induced by the vertices in $T$. The ***strength*** of edge $e$ is

$$\max \{\ K(G[U]) \ : \ e \subseteq U \subseteq V \ \}\,.$$

Note that $s_e$ is always at least $K(G)$, the min cut value in the original graph.

In the dumbbell example, the middle edge has strength 1 (any subgraph containing that edge cannot have min cut bigger than 1) and each clique edge has strength $n/2 - 1$ (take $U$ to be the vertices of that clique).

Strength can also be defined in the following equivalent way.

**Definition 4.2.** Define $(\ s_e \ : \ e \in E\ )$ to be the maximal values such that

$$\min \{\ u(\delta(U) \cap E_{s_e}) \ : \ e \in \delta(U)\ \} \ \ge \ s_e \qquad \text{where} \qquad E_x \ = \ \{\ f \in E \ : \ s_f \ge x\ \}\,. \tag{4.1}$$

The ***strength*** of edge $e$ is defined to be $s_e$.

**Theorem 4.3** (Benczúr-Karger [1]). Set

$$\begin{aligned} p_e \ &\ge \ u_e/s_e \qquad \forall e \in E \\ \rho \ &= \ L \log(n)/\epsilon^2. \end{aligned} \tag{4.2}$$

Then the weight vector $w$ output by Algorithm 2.1 satisfies the cut preservation condition (2.1) with high probability.

As we observed above, $s_e \ge K$ for all $e \in E$, so the hypothesis (4.2) is weaker than the hypothesis (3.1), so Theorem 4.3 is a strengthening of Theorem 3.1.

**Proof.** Using the definitions of $w_e$, $p_e$ and $E_x$, we have

$$w_e \ = \ \sum_{i \in [\rho]} \frac{u_e Z_{i,e}}{\rho p_e} \ = \ \frac{1}{\rho} \sum_{i \in [\rho]} \frac{u_e}{p_e s_e} Z_{i,e} \int_0^{s_e} dx \ = \ \frac{1}{\rho} \sum_{i \in [\rho]} \frac{u_e}{p_e s_e} Z_{i,e} \int_0^{\infty} 1_{e \in E_x}\, dx.$$

Thus

$$w(\delta(U)) \;=\; \frac{1}{\rho} \sum_{i\in[\rho],\, e\in\delta(U)} Z_{i,e} \frac{u_e}{p_e s_e} \int_0^\infty 1_{e\in E_x}\, dx \;=\; \int_0^\infty \underbrace{\frac{1}{\rho} \sum_{i\in[\rho],\, e\in\delta(U)\cap E_x} Z_{i,e} \frac{u_e}{p_e s_e}}_{=Y_{U,x}}\, dx. \tag{4.3}$$

We want to show that

$$w(\delta(U)) \;\in\; [1-\epsilon, 1+\epsilon] \cdot \mathrm{E}\,[\, w(\delta(U))\,] \qquad \forall U \subseteq V \tag{4.4}$$

with high probability. We will instead show that, with high probability, we have

$$Y_{U,x} \;\in\; [1-\epsilon, 1+\epsilon] \cdot \mathrm{E}\,[\, Y_{U,x}\,] \qquad \forall U \subseteq V,\, \forall x \geq 0. \tag{4.5}$$

This implies (4.4) via Claim 4.4. (Consider setting $f(x) = Y_{U,x}$ and $g(x) = (1+\epsilon)\,\mathrm{E}\,[\, Y_{U,x}\,]$.)

**Claim 4.4.** If $f(x) \leq g(x)$ for all $x \geq 0$ then $\int_0^\infty f(x)\,dx \leq \int_0^\infty g(x)\,dx$ (if both integrals exist).

The condition (4.5) seems difficult to prove because infinitely many constraints must hold simultaneously! However, things are not as bad as they seem: notice that there are at most $|E| \leq n^2$ distinct strength values (one for each edge), so the number of different sets $E_x$ is at most $n^2$. This means that, for each $U$, the number of different random variables $Y_{U,x}$ is at most $n^2$. So if we prove that, for each *fixed* $x$, that

$$Y_{U,x} \;\in\; [1-\epsilon, 1+\epsilon] \cdot \mathrm{E}\,[\, Y_{U,x}\,] \qquad \forall U \subseteq V \tag{4.6}$$

holds with probability at most $n^{-c}$, then a union bound implies that (4.5) holds with probability at most $n^{-c+2}$.

So fix any $x \geq 0$. We now analyze the probability that (4.6) holds. The amazing twist in the analysis is that we will prove it using Theorem 3.1! To do so, we create a new graph $(V, E_x)$ with edge weights $u'_e = u_e/s_e$ instead of $u_e$. Consider applying Algorithm 2.1 to this new graph with the same sampling probabilities $p_e$. What is the weight of the cut $\delta(U)$ in the sampled graph? It is

$$\sum_{e\in\delta(U)\cap E_x} \sum_{i=1}^{\rho} Z_{i,e} \frac{u'_e}{\rho p_e},$$

which is precisely $Y_{U,x}$ since we defined $u'_e = u_e/s_e$. So, if the sampled graph satisfies the cut preservation condition (2.1), then (4.6) holds simultaneously for all $U$. Theorem 3.1 will show that this holds with high probability, so long as we can show that the new graph satisfies the hypotheses of theorem. The remainder of the proof shows this.

Let $K$ refer to the minimum cut value in the new graph. We must show that $p_e \geq u'_e/K$. By (4.2) we have $p_e \geq u_e/s_e = u'_e$, so it suffices to show that $K \geq 1$. Consider any $U$ with $\delta(U) \cap E_x \neq \emptyset$. By definition of $s_e$, every $e \in \delta(U) \cap E_x$ has $s_e \leq u(\delta(U) \cap E_x)$. Thus

$$u'(\delta(U) \cap E_x) \;=\; \sum_{e\in\delta(U)\cap E_x} u_e/s_e \;\geq\; \sum_{e\in\delta(U)\cap E_x} u_e/u(\delta(U)\cap E_x) \;=\; 1.$$

As this holds for all $U$, we have $K \geq 1$. ∎

**Remarks**

- If the sampling probabilities are all $p_e = u_e/s_e$, then random sparsification decreases the number of edges to $O(n \log n/\epsilon^2)$ in expectation. This is because

$$\mathrm{E}\,[\,|\mathrm{supp}(w)|\,] \;=\; \sum_{e\in E} \Pr\,[\,w_e > 0\,] \;\leq\; \sum_{e\in E}\sum_{i\in[\rho]} \Pr\,[\,Z_{i,e} \neq 0\,] \;=\; \rho \sum_{e\in E} u_e/s_e,$$

  and it is shown by Benczúr and Karger that $\sum_{e\in E} u_e/s_e \leq n-1$.
- The strength values $s_e$ can be computed exactly in polynomial time, but they are somewhat unwieldy. Benczúr and Karger show how to compute in $O(n\,\mathrm{polylog}(n)/\epsilon^2)$ time sampling probabilities that satisfy (4.2) and still ensure that the number of edges is $O(n \log n/\epsilon^2)$ in expectation.

# 5 Matrix Concentration Bounds

## 5.1 Theorem Statement

Let $X$ be a random matrix of size $d \times d$. There are two different ways to think of a random matrix:

1. A matrix sampled according to a distribution on matrices
2. An array of scalar random variables

Our perspective also impacts how we interpret the expectation of a random matrix.

1. If we consider $X$ as sampled according to some distribution on matrices, then $\mathrm{E}[X] = \sum_A A \cdot \Pr[X = A]$.
2. If we consider $X$ as an array of random variables, then $\mathrm{E}[X]$ is the array of the expectations of the entries of $X$

Given independent, random, symmetric, positive semi-definite matrices $X_1, X_2, ..., X_k$, we want to understand the concentration of $\sum_i X_i$. Theorem 5.16 below is a recent result of Tropp [4] that solves this problem. In order to prove Tropp's theorem, we need to gather some definitions and results on symmetric matrices.

## 5.2 Löwner Ordering, Monotonicity, Convexity and Concavity

**Definition 5.1.** Let $A$ be any $d \times d$ symmetric matrix. The matrix $A$ is called **positive semi-definite** if all of its eigenvalues are non-negative. This is denoted $A \succeq 0$, where here 0 denotes the zero matrix. The matrix $A$ is called **positive definite** if all of its eigenvalues are strictly positive. This is denoted $A \succ 0$.

The positive semi-definite condition can be used to define a partial ordering on all symmetric matrices. This is called the **Löwner ordering** or the **positive semi-definite ordering**. For any two symmetric matrices $A$ and $B$, we write $A \succeq B$ if $A - B \succeq 0$.

For any $f : \mathbb{R} \to \mathbb{R}$, we can define a function on symmetric matrices $A$ by applying $f$ to the eigenvalues of $A$. Formally, let $A = UDU^T$ be the spectral decomposition of $A$. That is, $U$ is orthogonal and $D$ is the diagonal matrix whose diagonal entries are the eigenvalues of $A$.
Define $f(A) = Uf(D)U^T$, where $f(D)$ is a diagonal matrix with $[f(D)]_{ii} = f(D_{ii})$.

We will use this definition primarily with $f$ being exp or log.

**Claim 5.2.** Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ satisfy $f(x) \leq g(x)\ \forall x \in [l, u]$. Suppose $A$ is symmetric and the eigenvalues of $A$ all lie in $[l, u]$. Then $f(A) \preceq g(A)$.

How do functions behave with respect to the Löwner ordering? Usually badly. One might hope that if $f$ is monotone on some interval $[l, u]$, then when we extend $f$ to matrices, we obtain a monotone operator on matrices with eigenvalues in the interval $[l, u]$. Is it true that if $A \preceq B$ and the eigenvalues of $A, B$ are in $[l, u]$, then necessarily $f(A) \preceq f(B)$. Unfortunately not.

**Claim 5.3.** If $X$ and $Y$ are random matrices and $X \preceq Y$, then $\mathrm{E}[X] \preceq \mathrm{E}[Y]$.

While monotone funtions on $\mathbb{R}$ do not necessarily yield monotone functions on symmetric matrices as we saw above, it is true that if $f$ is monotone then $\operatorname{tr} f := A \mapsto \operatorname{tr}(f(A))$ is monotone. In order to establish this, we need a preliminary result concerning the spectrum of two matrices $A, B$ with $A \preceq B$.

**Claim 5.4** (Weyl's Monotonicity Theorem). Suppose $A$ and $B$ are symmetric, $n \times n$ matrices. Let $\lambda_i(A)$ be the $i$th largest eigenvalue of $A$. If $A \preceq B$, then $\lambda_i(A) \leq \lambda_i(B)$ for all $i$.

**Claim 5.5.** If $f$ is monotone, then $\operatorname{tr} f$ is monotone.

**Proof.** This follows easily from Claim 5.4. Say $A \preceq B$. We establish $\operatorname{tr} f(A) \leq \operatorname{tr} f(B)$:

$$\operatorname{tr} f(A) = \sum_{i=1}^{n} f(\lambda_i(A)) \leq \sum_{i=1}^{n} f(\lambda_i(B)) = \operatorname{tr} f(B)$$

■

We will use this result for $f = \exp$.

**Definition 5.6.** A function $f : \mathbb{R} \to \mathbb{R}$ is ***operator concave*** if

$$f((1-x)A + xB) \;\succeq\; (1-x)f(A) + xf(B) \qquad \forall x \in [0,1],\ \forall A, B.$$

Unfortunately, concavity of $f$ on $\mathbb{R}$ doesn't imply that $f$ is operator concave. However, the following claim is known.

**Claim 5.7.** log is operator concave.

Next, we define a new multiplication operation on positive definite matrices.

**Definition 5.8.** If $A, B$ are positive definite, define $A \odot B = \exp(\log(A) + \log(B))$.

This operation actually yields an Abelian group on the set of positive definite matrices. In particular, $\odot$ is commutative. Also, if $A$ and $B$ commute then $A \odot B$ is the usual product $AB$.

**Theorem 5.9.** (Lieb) Fix any symmetric $H$. The map $A \mapsto \operatorname{tr} \exp\big(\log(A) + H\big)$ is concave on positive definite matrices.

Lieb's theorem is difficult, and we will not be doing the proof.

**Corollary 5.10.** $\operatorname{tr}(A \odot B)$ is concave in $A$.

**Proof.** $\operatorname{tr}(A \odot B) = \operatorname{tr} \exp(\log A + \log B)$. Apply Lieb's theorem with $H = \log B$. ■

**Corollary 5.11.** Let $B$ be fixed, and $A$ a random matrix. Then $\mathrm{E}\,[\operatorname{tr}(A \odot B)] \leq \operatorname{tr}(\mathrm{E}\,[A] \odot B)$.

**Proof.** Apply Jensen's inequality. ■

**Corollary 5.12.** Let $A_1, ..., A_k$ be independent random positive definite matrices. Then

$$\mathrm{E}\,[\operatorname{tr}(A_1 \odot ... \odot A_k)] \;\leq\; \operatorname{tr}(\mathrm{E}\,[A_1] \odot ... \odot \mathrm{E}\,[A_k]).$$

**Proof.** Induction, applied to the preceding result. ■

## 5.3 The Chernoff Bound

**Theorem 5.13.** Let $X_1, ..., X_k$ be independent random variables with $0 \le X_i \le R$.
Let $\mu_{\min} \le \sum_i \mathrm{E}[X_i] \le \mu_{\max}$. Then, for all $\delta \ge 0$,

$$\Pr\left[\sum_{i=1}^k X_i \ge (1+\delta)\mu_{\max}\right] \overset{(a)}{\le} \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \overset{(b)}{\le} \begin{cases} e^{-\delta^2\mu_{\max}/3R} & (\text{if } \delta \le 1) \\ e^{-\delta\mu_{\max}/3R} & (\text{if } \delta > 1) \end{cases}$$

$$\Pr\left[\sum_{i=1}^k X_i \le (1-\delta)\mu_{\min}\right] \overset{(c)}{\le} \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \overset{(d)}{\le} e^{-\delta^2\mu_{\min}/2R}.$$

Inequalities (c) and (d) are only valid for $\delta \le 1$, but $\Pr\left[\sum_{i=1}^k X_i \le (1-\delta)\mu_{\min}\right] = 0$ if $\delta > 1$.

We now prove only inequality (a).

**Claim 5.14.**

$$\Pr\left[\sum_{i=1}^k X_i \ge t\right] \le \inf_{\theta>0} e^{-\theta t} \cdot \prod_{i=1}^k \mathrm{E}\left[e^{\theta X_i}\right].$$

**Proof.** Fix $\theta > 0$.

$$\begin{aligned} \Pr[\textstyle\sum_i X_i \ge t] &= \Pr[\textstyle\sum_i \theta X_i \ge \theta t] \\ &= \Pr[\exp(\textstyle\sum_i \theta X_i) \ge \exp(\theta t)] && \text{(monotonicity of } e^x\text{)} \\ &\le e^{-\theta t} \cdot \mathrm{E}[\exp(\textstyle\sum_i \theta X_i)] && \text{(Markov's inequality)} \end{aligned}$$

This expectation can be simplified:

$$\mathrm{E}[\exp(\textstyle\sum_i \theta X_i)] = \mathrm{E}\left[\prod_i e^{\theta X_i}\right] = \prod_i \mathrm{E}\left[e^{\theta X_i}\right] \qquad \text{(by independence).}$$

Combining these proves the claim. ■

**Claim 5.15.** Let $X$ be a random variable with $0 \le X \le 1$. Then

$$\mathrm{E}\left[e^{\theta X}\right] \le 1 + (e^\theta - 1) \cdot \mathrm{E}[X].$$

**Proof.** For $x \in [0,1]$ we have $e^{\theta x} \le 1 + (e^\theta - 1) \cdot x$, by convexity of the left-hand side. Since $X \in [0,1]$,

$$\begin{aligned} e^{\theta X} &\le 1 + (e^\theta - 1) \cdot X \\ \implies \qquad \mathrm{E}\left[e^{\theta X}\right] &\le 1 + (e^\theta - 1) \cdot \mathrm{E}[X], \end{aligned}$$

since inequalities are preserved under taking expectation. ■

**Proof** (of Chernoff Upper Bound). Without loss of generality $R = 1$.

$$\begin{aligned} \textstyle\prod_{i=1}^k \mathrm{E}\left[e^{\theta X_i}\right] &\le \textstyle\prod_{i=1}^k \left(1 + (e^\theta - 1) \cdot \mathrm{E}[X_i]\right) && \text{(by Claim 5.15)} \\ &= \exp\left(\textstyle\sum_{i=1}^k \log\left(1 + (e^\theta - 1) \cdot \mathrm{E}[X_i]\right)\right) \\ &\le \exp\left(\textstyle\sum_{i=1}^k (e^\theta - 1) \cdot \mathrm{E}[X_i]\right) && \text{(using } \log(1+x) \le x\text{)} \\ &\le \exp\left((e^\theta - 1)\mu_{\max}\right) \end{aligned}$$

Applying Claim 5.14 with $t = (1+\delta)\mu_{\max}$ and $\theta = \ln(1+\delta)$

$$\begin{aligned} \Pr\left[\sum_i X_i \ge (1+\delta)\mu_{\max}\right] &\le \exp\left(-\ln(1+\delta) \cdot (1+\delta)\mu_{\max}\right) \cdot \exp(\delta \cdot \mu_{\max}) \\ &= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}} \end{aligned}$$

■

## 5.4 Tropp's Matrix Chernoff Bound

**Theorem 5.16.** Let $X_1, ..., X_k$ be independent random $d \times d$ symmetric matrices with $0 \preceq X_i \preceq R \cdot I$. Let $\mu_{\min} \cdot I \preceq \sum_i \mathrm{E}[X_i] \preceq \mu_{\max} \cdot I$. Then, for all $\delta \in [0,1]$,

$$\begin{aligned}
\Pr\Big[\lambda_{\max}(\textstyle\sum_{i=1}^k X_i) \geq (1+\delta)\mu_{\max}\Big] &\overset{(a)}{\leq} d \cdot \Big(\tfrac{e^{\delta}}{(1+\delta)^{1+\delta}}\Big)^{\mu_{\max}/R} \overset{(b)}{\leq} d \cdot e^{-\delta^2 \mu_{\max}/3R} \\
\Pr\Big[\lambda_{\min}(\textstyle\sum_{i=1}^k X_i) \leq (1-\delta)\mu_{\min}\Big] &\overset{(c)}{\leq} d \cdot \Big(\tfrac{e^{-\delta}}{(1-\delta)^{1-\delta}}\Big)^{\mu_{\min}/R} \overset{(d)}{\leq} d \cdot e^{-\delta^2 \mu_{\min}/2R}.
\end{aligned}$$

Inequality (a) is actually valid for all $\delta \geq 0$.

We now prove inequality (a). Inequalities (b) and (d) follow from the discussion in the appendix.

**Claim 5.17.**

$$\Pr\left[\lambda_{\max}\Big(\sum_{i=1}^k X_i\Big) \geq t\right] \leq \inf_{\theta > 0} e^{-\theta t} \cdot \mathrm{tr}\left(\bigodot_{i=1}^k \mathrm{E}\left[e^{\theta X_i}\right]\right).$$

**Proof.** Fix $\theta > 0$.

$$\begin{aligned}
\Pr[\lambda_{\max}(\textstyle\sum_i X_i) \geq t] &= \Pr[\lambda_{\max}(\textstyle\sum_i \theta X_i) \geq \theta t] && \text{(homogeneity of max eigenvalue)} \\
&= \Pr\left[\exp\big(\lambda_{\max}(\textstyle\sum_i \theta X_i)\big) \geq \exp(\theta t)\right] && \text{(monotonocity of } e^x\text{)} \\
&\leq e^{-\theta t} \cdot \mathrm{E}\left[\exp\big(\lambda_{\max}(\textstyle\sum_i \theta X_i)\big)\right] && \text{(Markov's inequality)}
\end{aligned}$$

We can bound the maximum eigenvalue by a trace:

$$\begin{aligned}
\exp\big(\lambda_{\max}(\textstyle\sum_i \theta X_i)\big) &= \lambda_{\max}\big(\exp(\textstyle\sum_i \theta X_i)\big) && \text{(definition of matrix exponentiation)} \\
&\leq \mathrm{tr}\big(\exp(\textstyle\sum_i \theta X_i)\big) && \text{(max eigenvalue} \leq \text{sum of eigenvalues)}
\end{aligned}$$

Taking the expectation gives the bound:

$$\Pr[\lambda_{\max}(\textstyle\sum_i X_i) \geq t] \leq e^{-\theta t} \cdot \mathrm{E}\left[\mathrm{tr}\left(\exp(\textstyle\sum_i \theta X_i)\right)\right].$$

This expectation can be bounded:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{tr}\left(\exp(\textstyle\sum_i \theta X_i)\right)\right] &= \mathrm{E}\left[\mathrm{tr}\left(\exp(\textstyle\sum_i \log A_i)\right)\right] && \text{(let } A_i = e^{\theta X_i}\text{)} \\
&= \mathrm{E}\left[\mathrm{tr}(A_1 \odot \cdots \odot A_k)\right] && \text{(definition of } \odot\text{)} \\
&\leq \mathrm{tr}\left(\mathrm{E}[A_1] \odot \cdots \odot \mathrm{E}[A_k]\right) && \text{(by Corollary 5.12)}
\end{aligned}$$

Combining these inequalities proves the claim. ■

**Claim 5.18.** Let $X$ be a random symmetric $d \times d$ matrix with $0 \preceq X \preceq I$. Then

$$\mathrm{E}\left[e^{\theta X}\right] \preceq I + (e^{\theta} - 1) \cdot \mathrm{E}[X].$$

**Proof.** For $x \in [0,1]$ we have $e^{\theta x} \leq 1 + (e^{\theta} - 1) \cdot x$, by convexity of the left-hand side. Since $X$ has all eigenvalues in $[0,1]$, Claim 5.2 gives

$$\begin{aligned}
e^{\theta X} &\preceq I + (e^{\theta} - 1) \cdot X \\
\Longrightarrow \qquad \mathrm{E}\left[e^{\theta X}\right] &\preceq I + (e^{\theta} - 1) \cdot \mathrm{E}[X],
\end{aligned}$$

since the Löwner ordering is preserved under taking expectation by Claim 5.3. ■

**Proof** (of Matrix Chernoff Upper Bound). Without loss of generality $R = 1$. Our first observation is a bound for a sum of logs:

$$\begin{aligned}\textstyle\sum_{i=1}^k \log \mathrm{E}\left[\, e^{\theta X_i}\,\right] &= k \cdot \textstyle\sum_{i=1}^k \frac{1}{k} \log \mathrm{E}\left[\, e^{\theta X_i}\,\right] \\ &\preceq k \cdot \log\left(\textstyle\sum_{i=1}^k \frac{1}{k} \mathrm{E}\left[\, e^{\theta X_i}\,\right]\right) \quad \text{(by Claim 5.7)} \end{aligned} \tag{5.1}$$

Next:

$$\begin{aligned}&\mathrm{tr}\left(\mathrm{E}\left[\, e^{\theta X_1}\,\right] \odot \cdots \odot \mathrm{E}\left[\, e^{\theta X_k}\,\right]\right) \\ &\quad= \mathrm{tr}\exp\left(\textstyle\sum_{i=1}^k \log \mathrm{E}\left[\, e^{\theta X_i}\,\right]\right) && \text{(definition of } \odot) \\ &\quad\le \mathrm{tr}\exp\left(k \cdot \log\left(\textstyle\sum_{i=1}^k \frac{1}{k} \mathrm{E}\left[\, e^{\theta X_i}\,\right]\right)\right) && \text{(by (5.1) and Claim 5.5)} \\ &\quad\le d \cdot \lambda_{\max}\left(\exp\left(k \cdot \log\left(\textstyle\sum_{i=1}^k \frac{1}{k} \mathrm{E}\left[\, e^{\theta X_i}\,\right]\right)\right)\right) && \text{(sum of eigenvalues} \le d \text{ times maximum)} \\ &\quad\le d \cdot \exp\left(k \cdot \log \lambda_{\max}\left(\textstyle\sum_{i=1}^k \frac{1}{k} \mathrm{E}\left[\, e^{\theta X_i}\,\right]\right)\right) && \text{(definition of matrix exp and log)} \\ &\quad\le d \cdot \exp\left(k \cdot \log \lambda_{\max}\left(I + \textstyle\sum_{i=1}^k \frac{1}{k}(e^{\theta} - 1)\,\mathrm{E}\left[\, X_i\,\right]\right)\right) && \text{(by Claim 5.18)} \\ &\quad= d \cdot \exp\left(k \cdot \log\left(1 + \tfrac{e^{\theta}-1}{k} \lambda_{\max}\left(\textstyle\sum_{i=1}^k \mathrm{E}\left[\, X_i\,\right]\right)\right)\right) \\ &\quad\le d \cdot \exp\left((e^{\theta} - 1) \cdot \lambda_{\max}\left(\textstyle\sum_{i=1}^k \mathrm{E}\left[\, X_i\,\right]\right)\right) && \text{(using } \log(1+x) \le x) \\ &\quad\le d \cdot \exp\left((e^{\theta} - 1) \cdot \mu_{\max}\right)\end{aligned}$$

Apply Claim 5.17 with $t = (1+\delta)\mu_{\max}$ and $\theta = \ln(1+\delta)$:

$$\begin{aligned}\Pr\left[\, \lambda_{\max}\left(\textstyle\sum_i X_i\right) \ge (1+\delta)\mu_{\max}\,\right] &\le \exp\left(-\ln(1+\delta) \cdot (1+\delta)\mu_{\max}\right) \cdot \left(d \cdot \exp(\delta \cdot \mu_{\max})\right) \\ &= d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}}\end{aligned}$$

■

## References


[1] A. A. Benczúr and D. R. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs, 2002. `http://arxiv.org/abs/cs/0207078`.

[2] D. R. Karger. Random sampling in cut, flow, and network design problems. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing (STOC)*, 1994.

[3] D. R. Karger. Random sampling in cut, flow, and network design problems. *Mathematics of Operations Research*, 24(2):383–413, May 1999.

[4] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2011.