1 Motivation

One key ingredient to obtain a performance bound of spectral clustering is to estimate

$$L - \bar{L} = \sum_{i<j} (a_{ij} - E a_{ij})(e_i - e_j)(e_i - e_j)^\top.$$  

where $L$ is the Laplacian associated to the adjacency matrix and each simple rank-1 matrix is in the following form:

$$(e_i - e_j)(e_i - e_j)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$  

This is viewed as a sum of simple matrices with random coefficients.

Let’s recall in probability theory, we have see similar things: given a sequence of Bernoulli random variable $X_i$, what is its tail probability

$$ P \left( \frac{1}{n} \left| \sum_{i=1}^{n} (X_i - E X_i) \right| > t \right).$$  

We know that by law of large number, this quantity will go to 0 as $n \to \infty$ for any fixed $t$. However, here we are interested in the deviation of random matrix for its mean. Also, the law of large number is asymptotic: it does not specify the rate of probability decay w.r.t. $t$ and $t$.

In fact, the concentration of measure finds numerous applications in high dimensional statistical inference and geometric functional analysis. It is one of the cornerstones in modern data science. We will start from the concentration inequality for scalar random variables and then proceed to random matrix. This note provides a quick introduction to the vast literature of concentration inequalities. If you want to know more details, please refer to [1] and [3, Chapter 2].

2 Concentration inequality

To compare the tail bound of different concentration inequalities, we will use Bernoulli random variables as an example. We define $S_n := \sum_{i=1}^{n} X_i$ as the sum of $n$ i.i.d.
Bernoulli($p$) random variables and let $\mu := \mathbb{E} S_n = \mu$. We aim to estimate the following quantity:

$$P (|S_n - \mu| > \delta \mu).$$

The simplest method is to use moments to bound the tail probability. Most of you who have taken probability before know the famous Markov’s inequality.

**Theorem 2.1 (Markov’s inequality).** Let $X \geq 0$ be a **non-negative** random variable with $\mathbb{E}(X) < \infty$.

$$P(X \geq t) \leq \frac{\mathbb{E}(X)}{t}, \quad t > 0.$$  

Note that Markov’s inequality only holds for nonnegative random variables and assume a finite expectation.

**Proof:** Consider $Y := t \cdot 1_{\{X \geq t\}} \geq 0$. The following relation holds

$$Y = t \cdot 1_{\{X \geq t\}} \leq X \cdot 1_{\{X \geq t\}} \leq X.$$  

for any $X$ and $t > 0$. Next, we take its expectation:

$$E(Y) = E(t \cdot 1_{\{X \geq t\}}) = t E(1_{\{X \geq t\}} = t P(X \geq t) \leq E(X) \implies P(X \geq t) \leq \frac{E(X)}{t}.$$  

With Markov’s inequality, we are ready to provide the first bound:

$$P(S_n \geq (1 + \delta)\mu) \leq \frac{\mathbb{E} S_n}{(1 + \delta)\mu} = \frac{1}{1 + \delta}$$  

where $\mu = \mathbb{E} S_n$. Markov’s inequality provides a nontrivial bound if $\delta > 0$.

Despite its simplicity of Markov’s inequality, it is the starting point for many other inequalities. Here is a direction extension of Markov’s inequality by taking higher moments into consideration.

**Theorem 2.2 (Chebyshev’s inequality).** If $X$ has moments up to order $q \geq 1 \ i.e., \mathbb{E} |X|^q < \infty$. Then we have

$$P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^q)}{t^q}$$  

**Proof:** Let $Y = |X - \mathbb{E}(X)|^q \geq 0$ and apply Markov’s inequality to $Y$:

$$P(|X - \mathbb{E}(X)| \geq t) = P(|X - \mathbb{E}(X)|^q \geq t^q) = P(Y \geq t^q) \leq \frac{\mathbb{E} Y}{t^q}$$  

**Question:** Can you prove weak law of large number if each i.i.d. random variable $X_i$ satisfies $\mathbb{E} |X_i|^2 < \infty$?

Return to the example: the mean and variance of $S_n$ are

$$\mathbb{E} X = np, \quad \text{Var} S_n = np(1 - p).$$
Therefore, Chebyshev’s inequality implies
\[
P(|S_n - \mu| \geq \delta \mu) \leq \frac{\text{Var}(S_n)}{\delta^2 \mu^2} = \frac{np(1-p)}{\delta^2 n^2 p^2} = \frac{1-p}{\delta^2 np}.
\]
Compared with Markov’s inequality, we see that the tail bound depends on \(n\): as \(n \to \infty\), the probability decays to zero for any fixed \(\delta\) and \(p\). Apparently, we get a more refined bound by incorporating more information.

The next theorem is another version of Markov’s inequality and extremely useful when dealing with concentration of bounded, subexponential, and subgaussian random variables.

**Theorem 2.3** (Exponential version of Markov’s inequality). Let \(Y = \exp(\lambda X) > 0\) with \(\mathbb{E}(Y) < \infty\), then
\[
P(X \geq t) \leq \exp(-\lambda t) \mathbb{E}(\exp(\lambda X))
\]
where \(\lambda > 0\). In other words,
\[
P(X \geq t) \leq \inf_{\lambda \geq 0} \exp(-\lambda t) \mathbb{E}(\exp(\lambda X)).
\]
**Proof:** The proof follows from Markov’s inequality: note that for any positive \(\lambda\), it holds that
\[
P(X \geq t) = P(\lambda X \geq \lambda t) = P(\exp(\lambda X) \geq \exp(\lambda t)) \leq \exp(-\lambda t) \mathbb{E}(\exp(\lambda X)).
\]
where the last step uses Markov’s inequality.

**Question:** is the assumption \(\lambda > 0\) necessary? Which step uses \(\lambda > 0\)?

### 2.1 Chernoff’s inequality

To use the exponential version of Markov’s inequality, the key is to get an upper bound of \(\mathbb{E}(\exp(\lambda X))\). Next, we provide an example for \(X\) is a sum of independent Bernoulli random variables (not necessarily identical distributed).

**Theorem 2.4** (Chernoff’s inequality). Let \(\{X_i\}\) be a sequence of independent Bernoulli random variables with mean \(\{p_i\}\). Denote \(S_n = \sum_{i=1}^n X_i\) and \(\mu = \mathbb{E} S_n\), and for any \(\delta > 0\)
\[
P(S_n > (1 + \delta)\mu) < \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu = \exp(-\mu \left( (1 + \delta) \log(1 + \delta) - \delta \right))
\]
and
\[
P(S_n < (1 - \delta)\mu) < \left( \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\mu = \exp(-\mu \left( (1 - \delta) \log(1 - \delta) + \delta \right))
\]

**Remark 2.5.** The meaning of Chernoff’s inequality above does not seem obvious to us. In some applications, we would use the following slightly relaxed but more convenient version:
\[
P(S_n \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2 + \delta}\right), \quad \delta > 0
\]
\[
P(S_n \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right), \quad 0 < \delta < 1
\]
**Question:** Derive the relaxed Chernoff’s bound from the original version.

Example: For the sum of independent Bernoulli random variables:

\[ P(S_n \geq (1 + \delta)\mu) \leq \exp \left( -\frac{\delta^2 np}{2 + \delta} \right) \]

Compared with bound obtained from Markov’s and Chebyshev inequality, Chernoff’s bound provides a much tighter bound since the probability decay to zero exponentially fast w.r.t. \( n \).

**Proof of Chernoff’s inequality.** We first prove an important supporting result:

\[ \mathbb{E}\exp(\lambda X_i) \leq \exp(p_i(e^\lambda - 1)). \]

Following the definition of expectation, we have

\[ \mathbb{E}\exp(\lambda X_i) = \exp(\lambda) P(X_i = 1) + P(X_i = 0) = 1 + p_i(e^\lambda - 1). \]

Note that \( \exp(s) \geq 1 + s \) for any \( s \) since \( e^s \) is convex and \( 1 + s \) is the tangent of \( e^s \) at \( s = 0 \). Thus

\[ \mathbb{E}\exp(\lambda X_i) = 1 + p_i(e^\lambda - 1) \leq \exp(p_i(e^\lambda - 1)). \]

The main idea is to apply the exponential version of Markov’s inequality and the tricky part is to upper bound \( \mathbb{E}\exp(\lambda \sum_{i=1}^{n} X_i) \).

\[
\Pr \left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp(-\lambda t) \prod_{i=1}^{n} \mathbb{E}\exp(\lambda X_i) \quad \text{by independence}
\]

\[
\leq \exp(-\lambda t) \prod_{i=1}^{n} \exp(p_i(e^\lambda - 1)) \quad \text{use the supporting result}
\]

\[
= \exp(-\lambda t) \exp(\mu(e^\lambda - 1))
\]

where \( \mu = \sum_{i=1}^{n} p_i \). Now, letting \( t = (1 + \delta)\mu \), we have

\[
\Pr \left( \sum_{i=1}^{n} X_i \geq (1 + \delta)\mu \right) \leq \exp \left( -\lambda(1 + \delta)\mu + (e^\lambda - 1)\mu \right)
\]

Note that the bound is independent of \( \lambda \). Thus we can get a tighter bound by minimizing the right hand side over \( \lambda \). The minimum is achieved if

\[ e^\lambda = 1 + \delta \iff \lambda = \ln(1 + \delta). \]

Then

\[
\Pr \left( \sum_{i=1}^{n} X_i \geq (1 + \delta)\mu \right) \leq \exp \left( -\mu((1 + \delta)\ln(1 + \delta) - \delta) \right) = \frac{e^{\delta\mu}}{(1 + \delta)(1+\delta)^{\mu}}.
\]

\[ \square \]

Note that Chernoff’s inequality assumes \( X_i \) as a Bernoulli random variable. This assumption can be relaxed to bounded random variable, which leads to Hoeffding’s inequality.
2.2 Hoeffding’s inequality

In probability theory, Hoeffding’s inequality provides an upper bound on the tail probability of how much the sum of bounded independent random variables deviates from its expected value.

**Theorem 2.6** (Hoeffding’s inequality). Let \( X_i \) be a sequence of independent random variables with \( a_i \leq X_i \leq b_i \) with \( \mathbb{E}X_i = 0 \). Then

\[
\mathbb{P}\left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp\left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).
\]

Example: For Bernoulli random variables, we have \( X_i - \mathbb{E}X_i \in [-p, 1 + p] \).

\[
\mathbb{P}(S_n \geq (1 + \delta)\mu) = \mathbb{P}(S_n - \mu \geq \delta\mu) \leq \exp\left( -\frac{2\delta^2\mu^2}{n} \right) = \exp(-2\delta^2np^2)
\]

where \( t = \delta\mu \) and \( b_i - a_i = 1 \). Both bounds by Chernoff and Hoedding show exponential of tail probability w.r.t. \( n \). However, Hoedding is slightly more loose since it involves \( p^2 \) in the exponent instead of \( p \) in Chernoff’s bound.

The proof of Hoeffding’s inequality needs the following key lemma.

**Lemma 2.7** (Hoeffding’s Lemma). If \( a \leq X \leq b \) and \( \mathbb{E}(X) = 0 \), then

\[
\mathbb{E}(\exp(\lambda X)) \leq \exp\left( \frac{\lambda^2(b - a)^2}{8} \right).
\]

We don’t provide the proof here; you may find it in [1]. Note that the right hand side depends on \( \lambda^2 \) instead of \( \lambda \). Let’s try a special case: if we let \( X = X_i - p \) where \( X_i \) is Bernoulli(\( p \)), then

\[
\mathbb{E}\exp(\lambda(X_i - p)) \leq \exp\left( \frac{\lambda^2}{8} \right) \iff \mathbb{E}\exp(\lambda X_i) \leq \exp\left( p\lambda + \frac{\lambda^2}{8} \right)
\]

and on the other hand,

\[
\mathbb{E}\exp \lambda X_i \leq \exp(p(e^\lambda - 1)).
\]

**Proof of Hoeffding’s inequality.** The proof follows from Markov’s inequality and Hoeffding’s Lemma.

\[
\mathbb{P}\left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp(-\lambda t) \mathbb{E}\exp\left( \lambda \sum_{i=1}^{n} X_i \right)
\]

\[
= \exp(-\lambda t) \prod_{i=1}^{n} \mathbb{E}\exp(\lambda X_i) \quad \text{by independence}
\]

\[
= e^{\lambda t} \prod_{i=1}^{n} \exp\left( \frac{\lambda^2(b_i - a_i)^2}{8} \right) \quad \text{by Hoeffding’s Lemma}
\]

\[
= \exp\left( -\lambda t + \frac{\lambda^2 \sum_{i=1}^{n} (b_i - a_i)^2}{8} \right)
\]

We minimize the exponent over \( \lambda \) to get a tighter bound:

\[
\lambda^* = \frac{4t}{\sum_{i=1}^{n} (b_i - a_i)^2}
\]
Then
\[ P \left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp \left( -\frac{t^2}{4 \cdot \sum_{i=1}^{n} (b_i - a_i)^2} \right) = \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right). \]

\[ \square \]

2.3 Bernstein inequality

Now we present the last inequality: Bernstein’s inequality, which takes the variance into consideration.

**Theorem 2.8** (Bernstein’s inequality). Let \( X_i \) be a sequence of centered independent random variables:

\[ \mathbb{E}(X_i) = 0, \quad |X_i| \leq R, \quad \sigma^2 = \sum_{i=1}^{n} \mathbb{E} X_i^2. \]

Then
\[ P \left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp \left( -\frac{t^2/2}{Rt/3 + \sigma^2} \right) \]

**Example:** For \( S_n = \sum_{i=1}^{n} X_i, X_i \sim \text{Bernoulli}(p), p < 1/2 \)

\[ |X_i - \mathbb{E}(X_i)| = |X_i - p| \leq 1 - p \]

and
\[ \sigma^2 = \sum_{i=1}^{n} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 = np(1-p). \]

Thus letting \( \mu = \mathbb{E} S_n = np \), we have
\[ P (S_n - \mu \geq \delta \mu) \leq \exp \left( -\frac{\delta^2 \mu^2}{\delta \mu(1-p)/3 + np(1-p)} \right) \]

\[ = \exp \left( -\frac{\delta^2 n^2 p^2}{\delta np(1-p)/3 + np(1-p)} \right) \]

\[ \leq \exp \left( -\frac{\delta^2 np}{(1-p)(1+\delta/3)} \right) \]

If \( \delta \) is large, then
\[ P (S_n - \mu \geq \delta \mu) \leq \exp \left( -\Omega \left( \frac{\delta np}{1-p} \right) \right) \]

and if \( \delta \) is small,
\[ P (S_n - \mu \geq \delta \mu) \leq \exp \left( -\Omega \left( \frac{\delta^2 np}{1-p} \right) \right) \]

For the sum of Bernoulli random variables, Bernstein’s inequality provides a similar tail bound as Chernoff’s bound.

**Proof:** Still, we will apply Markov’s inequality. The key is to bound \( \mathbb{E} e^{\lambda X_i} \) since
\[ P(S_n \geq t) \leq e^{-\lambda t} \mathbb{E} \exp \left( \lambda \sum_{i=1}^{n} X_i \right) \]
\[ = e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E} \exp(\lambda X_i) \]
by independence.

The estimation of $\mathbb{E}\exp(\lambda X_i)$ relies on Taylor’s expansion:

$$\exp(\lambda X_i) = \sum_{k=0}^{\infty} \frac{\lambda^k X^k_i}{k!} = 1 + \lambda X_i + \frac{\lambda^2 X^2_i}{2} + \cdots$$

where

$$\exp(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!}$$

Now we take its expectation:

$$\mathbb{E}\exp(\lambda X_i) = \mathbb{E}\sum_{k=0}^{\infty} \frac{\lambda^k X^k_i}{k!} = 1 + \frac{\lambda^2 \sigma^2_i}{2} + \sum_{k\geq 3} \frac{\lambda^k}{k!} \mathbb{E} X^k_i$$

where $\mathbb{E} X_i = 0$.

Note that $X^k_i \leq |X_i|^k \leq R^{k-2} |X_i|^2$

$$|\mathbb{E} X^k_i| \leq \mathbb{E} |X_i|^k = \mathbb{E} |X_i|^{k-2} |X_i|^2 \leq R^{k-2} \mathbb{E} X^2_i = R^{k-2} \sigma^2_i$$

where $|X_i| \leq R$. Then

$$\mathbb{E} e^{\lambda X_i} \leq 1 + \frac{\lambda^2 \sigma^2_i}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k R^{k-2}}{k!} \sigma^2_i$$

For the third term,

$$\sum_{k=3}^{\infty} \frac{\lambda^k R^{k-2}}{k!} \sigma^2_i = \frac{\sigma^2_i}{R^2} \sum_{k=3}^{\infty} \frac{\lambda^k R^k}{k!} = \frac{\sigma^2_i}{R^2} \left( \sum_{k=0}^{\infty} \frac{\lambda^k R^k}{k!} - 1 - \lambda R - \frac{\lambda^2 R^2}{2} \right) = \frac{\sigma^2_i}{R^2} \left( e^{\lambda R} - 1 - \lambda R - \frac{\lambda^2 R^2}{2} \right)$$

where $e^{\lambda R} = 1 + \lambda R + \frac{\lambda^2 R^2}{2} + \cdots$

$$\mathbb{E} e^{\lambda X_i} \leq 1 + \frac{\lambda^2 \sigma^2_i}{2} + \frac{\sigma^2_i}{R^2} \left( e^{\lambda R} - 1 - \lambda R - \frac{\lambda^2 R^2}{2} \right)$$

$$= \frac{\sigma^2_i}{R^2} \left( e^{\lambda R} - 1 - \lambda R \right) + 1 \leq \exp \left( \frac{\sigma^2_i}{R^2} (e^{\lambda R} - 1 - \lambda R) \right)$$

Here we use $e^s \geq 1 + s$ again.

As a result, we have

$$\mathbb{P} \left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp(-\lambda t) \mathbb{E} \exp \left( \lambda \sum_{i=1}^{n} X_i \right)$$

$$= \exp(-\lambda t) \prod_{i=1}^{n} \mathbb{E} \exp(\lambda X_i)$$

$$\leq \exp(-\lambda t) \prod_{i=1}^{n} \exp \left( \frac{\sigma^2_i}{R^2} (e^{\lambda R} - 1 - \lambda R) \right)$$

$$= \exp \left( \frac{\sigma^2}{R^2} (e^{\lambda R} - 1 - \lambda R) - \lambda t \right)$$

7
Take the derivative w.r.t. $\lambda$:

$$\frac{\sigma^2}{R} e^{\lambda R} - \frac{\sigma^2}{R} - t = 0 \implies \lambda = \frac{1}{R} \log \left( 1 + \frac{Rt}{\sigma^2} \right)$$

The probability is bounded by

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

$$= \frac{\sigma^2}{R^2} \left( 1 + \frac{Rt}{\sigma^2} - 1 \right) - \frac{1}{R} \left( \frac{\sigma^2}{R} + t \right) \log \left( 1 + \frac{Rt}{\sigma^2} \right)$$

$$= \frac{t}{R} - \frac{1}{R} \left( \frac{\sigma^2}{R} + t \right) \log \left( 1 + \frac{Rt}{\sigma^2} \right)$$

$$= \frac{\sigma^2}{R^2} \cdot \frac{Rt}{\sigma^2} - \frac{\sigma^2}{R^2} \left( 1 + \frac{Rt}{\sigma^2} \right) \log \left( 1 + \frac{Rt}{\sigma^2} \right) \quad \text{by } u = \frac{Rt}{\sigma^2}$$

$$= \frac{\sigma^2}{R^2} (u - (1 + u) \log(1 + u)).$$

It can be shown that

$$u - (1 + u) \log(1 + u) \leq -\frac{u^2}{2 + 2u/3}, \quad u \geq 0.$$

Using the fact finishes the proof.

References

