Chapter 15
Some applications of concentration inequalities

15.1 Some large-scale matrix computation with randomization

Next is an application of the concentration inequality to a few large-scale computational problems in scientific computing. For concreteness, let us consider the problem of estimating the trace of a large matrix $Tr(\mathscr{A})$ (other interesting problems can be found in [4]). From Lemma 12.2 we know that

$$Tr(\mathscr{A}) = \mathbb{E}\left[\mathbf{m}^T \mathscr{A} \mathbf{m}\right] \overset{\text{Monte Carlo}}{\approx} \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}_i^T \mathscr{A} \mathbf{m}_i =: S_N,$$

where $\mathbf{m}$ is an arbitrary random vector with zero mean and $\mathbb{E}\left[\mathbf{m}\mathbf{m}^T\right] = I$. Clearly, S_N is an unbiased estimator for $Tr(\mathscr{A})$ and it converges almost surely to $Tr(\mathscr{A})$. The question that we are interested here is how many samples is "enough" for such an estimator. Addressing this question is of practical important since we want to minimize the cost, i.e. choosing as small ensemble size N as we can, while having an accurate estimator. In this section we consider symmetric positive definite matrix $\mathscr{A}$ which admits computable lower and upper bounds [4] for $\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i$, i.e.,

$$L_i \le \mathbf{m}_i^T \mathscr{A} \mathbf{m}_i \le U_i.$$

Thus each $\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i$ is a bounded random variable with mean $\mathbb{E}\left[\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i\right] = Tr(\mathscr{A})$. Applying the Hoeffding inequality from Exercise 14.3 we have

$$\mathbb{P}\left[|S_N - Tr(\mathscr{A})| > t\right] \le 2e^{-2N^2 \frac{t^2}{\Sigma_{i=1}^N (U_i - L_i)^2}}, \quad \forall t \ge 0.$$

In other words,

$$-t \le S_N - Tr(\mathscr{A}) \le t \tag{15.1}$$

holds with probability

$$1 - 2e^{-2N^2 \frac{t^2}{\Sigma_{i=1}^N (U_i - L_i)^2}}.$$

If we now can pick a tolerance t and a success probability β, we can solve

$$1-2e^{-2N^2\frac{t^2}{\sum_{i=1}^N (U_i-L_i)^2}}=\beta$$

for the ensemble size N as

$$N=\sqrt{\frac{\sum_{i=1}^N (U_i-L_i)^2}{2t^2}\ln\left(\frac{2}{1-\beta}\right)},$$

which is the desire ensemble size to obtain the error bound (15.1) with probability β.

Exercise 15.1. Your task is to estimate the trace of the matrix in the first numerical example of [4]. Pick a few pairs (t,β) (some with big β, say, 0.5) and report your observations.

15.2 Dimension reduction with random projection

We now show that concentration inequalities are the key to the success of many randomized methods for dimension reduction of big data. To begin, assume $\mathbf{x}\in\mathbb{R}^N$ and consider a random matrix $\mathscr{A}\in\mathbb{R}^{n\times N}$ whose entries, $\mathscr{A}_{ij}$ *are i.i.d random variables with zero mean and unit variance*. Let us define the following random "projection" $\mathscr{P}$

$$\mathbf{z}=\mathscr{P}\mathbf{x}:=\frac{1}{\sqrt{n}}\mathscr{A}\mathbf{x},$$

and we are going to show that with high probability the *random projection* $\mathscr{P}$ preserves length.

Exercise 15.2. Show that components of $\mathbf{z}$, i.e. $\mathbf{z}_i=\mathscr{A}(i,:)\mathbf{x}, i=1\ldots,n$ are i.i.d. random variables with

$$\mathbb{E}[\mathbf{z}_i]=0,\quad \text{and } \mathbb{V}ar[\mathbf{z}_i]=\frac{\|\mathbf{x}\|^2}{n},$$

In particular, show that $\mathbb{E}\left[\|\mathbf{z}\|^2\right]=\|\mathbf{x}\|^2$.

Exercise 15.2 shows a remarkable fact that, on average, mapping via random matrix with i.i.d. random entries having mean zero and unit variance preserves the length. Clearly, we are more interested in the performance of an actual realization of $\mathbf{z}$, i.e., how far $\mathbf{z}^2$ is from its mean. We address this question probabilistically. To that end we observe that $\mathbf{z}_i^2$ are i.i.d. random variables and the expectation of their sum is $\|\mathbf{x}\|^2$, and we shall show that the sum actually concentrates around the mean. By Chernoff inequality we have

$$\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon \|\mathbf{x}\|^2\right] = \mathbb{P}\left[n\|\mathbf{z}\|^2 \geq n(1+\varepsilon)\|\mathbf{x}\|^2\right]$$
$$\leq \min_{\lambda} e^{-n\lambda(1+\varepsilon)\|\mathbf{x}\|^2} \Pi_{i=1}^{n} \mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right],$$

and similar to other concentration results that we have seen so far, the task at hand is to bound the MGF $\mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right]$. Before considering general sub-gaussian random variable, let us first study the case when $\mathscr{A}_{ij}$ are standard normal random variables, for which we can evaluate the MGF exactly.

Exercise 15.3. Suppose $\zeta \sim \mathscr{N}\left(0, \sigma^2\right)$ and $t \leq 1/\left(2\sigma^2\right)$. Show that

$$\mathbb{E}\left[e^{t\zeta^2}\right] = \frac{1}{\sqrt{1-2t\sigma^2}}.$$

Then deduce that $\mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right] = \dfrac{1}{\sqrt{1-2\lambda\|\mathbf{x}\|^2}}$ for $\lambda \leq 1/\left(2\|\mathbf{x}\|^2\right)$.

Hint. Direct evaluation of the integral and the fact that $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} e^{-\frac{\theta^2}{2}} d\theta = 1$.

The result of Exercise 15.3 together with the fact that $\mathbf{z}_i$ are i.i.d. yields *Check it*

$$\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon \|\mathbf{x}\|^2\right] \leq \min_{\lambda} e^{\frac{n}{2} f(\lambda)},$$

where we have defined $f(\lambda) := -2\lambda(1+\varepsilon)\|\mathbf{x}\|^2 - \ln\left(1-2\lambda\|\mathbf{x}\|^2\right)$. It is easy to see that, for $0 \leq \lambda \leq 1/\left(2\|\mathbf{x}\|^2\right)$, $f(\lambda)$ attains its minimum at $\lambda^* = \dfrac{\varepsilon}{2(1+\varepsilon)\|\mathbf{x}\|^2}$. *Show it*

Thus, we have

$$\mathbb{P}\left[\|\mathbf{z}\|^2 \geq (1+\varepsilon)\|\mathbf{x}\|^2\right] \leq e^{\frac{n}{2}[\ln(1+\varepsilon)-\varepsilon]} \leq e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)},$$

where we have used the fact that, for $\varepsilon \in [0,1]$, $\ln(1+\varepsilon) - \varepsilon \leq -\dfrac{\varepsilon^2}{2} + \dfrac{\varepsilon^3}{3}$. *Check it*

Exercise 15.4. Show that

$$\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\|\mathbf{x}\|^2\right] \leq e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.$$

Together with the union bound we obtain the following concentration inequality

$$\boxed{\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\|\mathbf{x}\|^2 \text{ or } \|\mathbf{z}\|^2 \geq (1+\varepsilon)\|\mathbf{x}\|^2\right] \leq 2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.} \tag{15.2}$$

That is, ε-distortion in length via the random projection $\mathscr{P}$ is less than $2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}$. In other words, with high probability, i.e., $1 - 2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}$, the random projection $\mathscr{P}$ preserves the length of $\mathbf{x}$.

Exercise 15.5 (Concentration inequality for χ^2 distribution). We have essentially proved a concentration inequality for Chi-square distribution. To see this, let us recall a Chi-square distribution with n degrees of freedom, typically denoted as χ_n^2, is the sum of square of n i.i.d. standard normal random variables. That is, if $m \sim \chi_n^2$, then X can be represented as $m = \sum_{i=1}^{n} \xi_k^2$ where $\xi_k \sim \mathcal{N}(0,1)$. Show that m concentrates around its mean with the tail bound given by (15.2).

Hint. Rescale the random projection matrix $\mathscr{P}$ appropriately and judiciously pick a vector $\mathbf{x}$ so that $\|\mathbf{z}\|^2 \sim \chi_n^2$.

Suppose now we have m vectors $\mathbf{x}_i \in \mathbb{R}^N, i = 1,\dots,m$, where $N \gg 1$ and we are interested in reducing the dimension of these vectors while (approximately) preserving their geometry. We are going to show that this task can be acommplished by the random projection $\mathbf{y}_i = \mathscr{P}\mathbf{x}_i \in \mathbb{R}^n$. Here, by preserving geometry we mean

$$\|\mathbf{y}_i\| \approx \|\mathbf{x}_i\|, \quad \text{and } \|\mathbf{y}_i - \mathbf{y}_j\| \approx \|\mathbf{x}_i - \mathbf{x}_j\|,$$

that is we like to reduce the dimension of the data vectors, but we desire to preserve their norm and their mutual distances. Since $\mathscr{P}$ is linear, it is sufficient to show that the latter holds. Since we have m vectors, we have $m(m-1)/2$ distinct difference vectors $\mathbf{x}_i - \mathbf{x}_j$ (called "pairs"). If we project each "pair" $\mathbf{x}_i - \mathbf{x}_j$ to obtain the corresponding vector $\mathbf{y}_i - \mathbf{y}_j$, the union bound and the concentration inequality (15.2) give

$$\begin{aligned}\mathbb{P}[\text{Some pair has } \varepsilon\text{-distortion}] &\le \sum_{i=1}^{m(m-1)/2} \mathbb{P}[\text{Pair } i \text{ has } \varepsilon\text{-distortion}] \\ &= \frac{m(m-1)}{2}\mathbb{P}[\text{Pair 1 has } \varepsilon\text{-distortion}] \le m(m-1)\, e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.\end{aligned}$$

Thus, if desire

$$\mathbb{P}[\text{Some pair has } \varepsilon\text{-distortion}] \le \frac{1}{m^\beta},$$

we can enforce

$$m(m-1)\, e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)} \le \frac{1}{m^\beta},$$

i.e.,

$$\boxed{n \ge \frac{2\beta\log(m) + 2\log(m(m-1))}{\varepsilon^2/2 - \varepsilon^3/3}.} \tag{15.3}$$

We have derived a version of the Johnson-Lindenstrauss lemma.

Lemma 15.1 (Johnson-Lindenstrauss Lemma). *Consider m vectors $\mathbf{x}_i \in \mathbb{R}^N, i = 1,\dots,m$, where $N \gg 1$. Define a random projection matrix $\mathscr{P} = \mathscr{A}/\sqrt{n} \in \mathbb{R}^{n\times N}$, where each component of $\mathscr{A}$ is i.i.d. standard normal random variable. For any $\beta > 0$, if we choose a reduced dimension n satisfying* (15.3)*, then with probability at least $1 - m^{-\beta}$ we have*

$$(1-\varepsilon)\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2 \le \left\|\mathscr{P}\mathbf{x}_i-\mathscr{P}\mathbf{x}_j\right\|^2 \le (1+\varepsilon)\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2, \quad \forall i,j=1,\ldots,m.$$

Remark 15.1. Lemma 15.1 shows that i.i.d. Gaussian random matrices can be used to reduce the dimension of high-dimensional data sets while almost preserving the geometry with high probability. The beauty here is that the reduced dimension is independent of the original data dimension. Unlike other dimensional reduction methods, random projections do not require the low dimensionality of the original data set.

Exercise 15.6. In this exercise, we are going to numerically verify the Johnson-Lindenstrauss lemma. To that end, pick a few dimensions for the original *big-data*, say $N=\{100,1000,3000,6000\}$. For each N, generate $m=3000$ uniform (or Gaussian) random vectors of dimensional N. Pick a distortion value, say $\varepsilon=0.25$, and β so that the successful probability is 0.75. Now pick the minimum reduced dimension n using (15.3) and then compute the actual distortions for m data vectors and plot the histogram of the actual distortions for each N. Plot the mean, the min, and the max actual distortions for each N and, on the same figure for each case, plot the predicted distortion lines $1+\varepsilon$ and $1-\varepsilon$. Discuss on the validity of the Johnson-Lindenstrauss lemma. You may want check various values of β ranging from low to high successful probability. Also, pick a few values of ε and report your findings.

Let us now extend the result to sub-gaussian distributions, i.e., $\mathscr{A}_{ij}$ are i.i.d. sub-gaussian random variables with zero mean and unit variance. All we need is to bound the MGF $\mathbb{E}\left[e^{\lambda n\mathbf{z}_i^2}\right]$ which, unlike the Gaussian case, does not admit a closed form expression. Nevertheless, by Proposition 13.2, $\mathbf{z}_i$ is a sub-gaussian with proxy $\alpha^2\left\|\mathbf{x}\right\|^2/n$ and $n\mathbf{z}_i^2-\left\|\mathbf{x}^2\right\|$ is a zero-mean random variable. We will use this fact to bound the tail and in turn the MGF of $n\mathbf{z}_i^2-\left\|\mathbf{x}\right\|^2$. We start with the following general result.

Lemma 15.2. *Let m be a zero-mean random variable with the tail bound*

$$\mathbb{P}\left[|m|\ge t\right] \le 2e^{-2t/\beta} \tag{15.4}$$

for some $\beta>0$. Then the MGF of m satisfies

$$\mathbb{E}\left[e^{sm}\right] \le e^{2s^2\beta^2}, \quad \forall |s| \le \frac{1}{2\beta}. \tag{15.5}$$

Proof. We observe that the statement is similar to the equivalent between *i*) and *iv*) in Theorem 13.1. Indeed, similar to the proof of *ii*) in Theorem 13.1 we have

$$\mathbb{E}\left[|m|^p\right] = 2\left(\beta/2\right)^p p\Gamma\left(p\right) \le \beta^p p!$$

which, together with the fact that $\mathbb{E}\left[m\right]=0$, gives

$$\mathbb{E}\left[e^{sm}\right] = 1+\sum_{p=2}^{\infty}\frac{s^p\mathbb{E}\left[m^p\right]}{p!} \le 1+\sum_{p=2}^{\infty}\left(s\beta\right)^p = 1+\frac{s^2\beta^2}{1-s\beta},$$

for $|s\beta| < 1$, and for $|s|\,\beta \leq 1/2$ we arrive at

$$\mathbb{E}\left[e^{sm}\right] \leq 1 + 2s^2\beta^2 \leq e^{s^2\beta^2},$$

which ends the proof.

To apply Lemma 15.2 to the random variable $n\mathbf{z}_i^2 - \|\mathbf{x}\|^2$ we just need to bound its tail. We have

$$\mathbb{P}\left[n\mathbf{z}_i^2 - \|\mathbf{x}\|^2 > \varepsilon\,\|\mathbf{x}\|^2\right] = \mathbb{P}\left[|\mathbf{z}_i| > \sqrt{\frac{1+\varepsilon}{n}}\,\|\mathbf{x}\|\right] \leq 2e^{-\frac{(1+\varepsilon)\|\mathbf{x}\|^2}{2\alpha^2\|\mathbf{x}\|^2}} \leq 2e^{-\frac{\varepsilon\|\mathbf{x}\|^2}{2\alpha^2\|\mathbf{x}\|^2}},$$

where have used the tail bound of $\mathbf{z}_i$ in the first inequality. Now applying Lemma 15.2 with $t = \varepsilon\,\|\mathbf{x}\|^2$ and $\beta = 4\alpha^2\,\|\mathbf{x}\|^2$ we have

$$\mathbb{E}\left[e^{\lambda\left(n\mathbf{z}_i^2 - \|\mathbf{x}\|^2\right)}\right] \leq e^{32\alpha^4\|\mathbf{x}\|^4\lambda^2}.$$

Consequently, for $\varepsilon < 8\alpha^2$, we have

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon\,\|\mathbf{x}\|^2\right] &\leq \min_{\lambda} e^{-n\varepsilon\lambda\|\mathbf{x}\|^2}\Pi_{i=1}^n\mathbb{E}\left[e^{\lambda\left(n\mathbf{z}_i^2 - \|\mathbf{x}\|^2\right)}\right] \\ &\leq \min_{\lambda} e^{32\alpha^4\|\mathbf{x}\|^4\lambda^2 - \varepsilon\lambda\|\mathbf{x}\|^2} = e^{-n\frac{\varepsilon^2}{128\alpha^4}},\end{aligned}$$

and together with the union bound we obtain the concentration inequality

$$\boxed{\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\,\|\mathbf{x}\|^2 \text{ or } \|\mathbf{z}\|^2 \geq (1+\varepsilon)\,\|\mathbf{x}\|^2\right] \leq 2e^{-n\frac{\varepsilon^2}{128\alpha^4}}.}$$

Similar to the Gaussian case, we conclude that with high probability, i.e., $1 - 2e^{-n\frac{\varepsilon^2}{128\alpha^4}}$, the random projection $\mathscr{P}$ with i.i.d α-sub-gaussian random entries preserves the length of $\mathbf{x}$. This shows that the Johnson-Lindenstrass Lemma is also valid for sub-gausian random projection.

Note that random variables with MGF satisfying (15.5) is called *sub-exponential*. The name is due to the exponential tail (15.4), which is heavier than Gaussian tail of sub-gaussian random variables. As a by-product of proving the Johnson-Lindenstrauss lemma for sub-gaussian distrubtions, we have shown that square of sub-gaussian random variables is sub-exponential[1]. More importantly, we have shown that sub-exponential distributions also obey concentration inequalities. A more complete concentration inequality for sub-exponential random variables is given by Bernstein's inequality.

Theorem 15.1 (Bernstein's inequality). *Let $m_i, i = 1,\ldots,N$ be zero mean sub-exponential random variables, i.e., m_i satisfies the tail bound*

[1] In fact product of two sub-gaussian distributions is sub-exponential.

$$\mathbb{P}\left[|m_i| \geq t\right] \leq 2e^{-2t/\beta_i}.$$

Then, for any $t \geq 0$, we have

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} m_i\right| \geq t\right] \leq 2\min\left\{e^{-\frac{t^2}{8\Sigma_{i=1}^{N}\beta_i^2}}, ce^{-\frac{t}{2\max_i \beta_i}}\right\},$$

for some constant $c = e^{\frac{\Sigma_{i=1}^{N}\beta_i^2}{2\max_i \beta_i^2}}$.

Proof. Again, Chernoff inequality is the key of the proof. We have

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq t\right] \leq \min_{\lambda>0} \frac{\Pi_{i=1}^{N}\mathbb{E}\left[e^{\lambda m_i}\right]}{e^{\lambda t}} \leq \min_{\lambda \leq 1/(2\max_i \beta_i)} e^{2\lambda^2 \Sigma_{i=1}^{N}\beta_i^2 - \lambda t},$$

where we have used (15.5). Now optimizing λ, i.e. $\lambda^* = \min\left\{\frac{t}{4\Sigma_{i=1}^{N}\beta_i^2}, \frac{1}{2\max_i \beta_i}\right\}$, we obtain

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq t\right] \leq \min\left\{e^{-\frac{t^2}{8\Sigma_{i=1}^{N}\beta_i^2}}, ce^{-\frac{t}{2\max_i \beta_i}}\right\},$$

and this concludes the proof.

Compared to Proposition 13.2, the tail bound for sum of independent sub-exponential random variables is only Gaussian for small deviation t. For large deviation, the tail is exponential, and hence is heavier than Gaussian tails. Analogous results to Theorem 14.1 and Corollary 14.1 are given in the following exercises.

Exercise 15.7. Let $m_i, i = 1, \ldots, N$ be sub-exponential random variables as in Theorem 15.1 and $\mathbf{a} \in \mathbb{R}^N$. Show that

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} \mathbf{a}_i m_i\right| \geq t\right] \leq 2\min\left\{e^{-\frac{t^2}{c_1}}, c_2 e^{-\frac{t}{c_3}}\right\},$$

and determine c_1, c_2, c_3. Then deduce the following upper bound

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} m_i\right| \geq t\right] \leq 2\min\left\{e^{-N\frac{t^2}{c_1}}, c_2 e^{-N\frac{t}{c_3}}\right\}, \tag{15.6}$$