Lecture 15: Matrix deviation inequality

May 29 - June 3, 2020
A is an $m \times n$ random matrix whose rows are independent, mean zero, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. 

(If you find it helpful to think in terms of concrete examples, let the entries of $A$ be independent $\mathcal{N}(0, 1)$ random variables.)

For a fixed vector $x \in \mathbb{R}^n$, we have

\[
\mathbb{E}\|Ax\|_2^2 = \mathbb{E} \sum_{j=1}^{m} (A_{j,:}x)^2 = \sum_{j=1}^{m} \mathbb{E}(A_{j,:}x)^2 \\
= \sum_{j=1}^{m} x^T \mathbb{E}(A_{j,:}^TA_{j,:})x = m\|x\|_2^2.
\]

Further, if we assume that concentration about the mean holds here (and in fact, it does), we should expect that

\[
\|Ax\|_2 \approx \sqrt{m}\|x\|_2
\]

with high probability.
Similarly to Johnson-Lindenstrauss Lemma, our next goal is to make the last formula hold simultaneously over all vectors \( \mathbf{x} \) in some fixed set \( \mathcal{T} \subset \mathbb{R}^n \). Precisely, we may ask – how large is the average uniform deviation:

\[
\mathbb{E} \sup_{\mathbf{x} \in \mathcal{T}} \left| \| A\mathbf{x} \|_2 - \sqrt{m} \| \mathbf{x} \|_2 \right|
\]

This quantity should clearly depend on some notion of the size of \( \mathcal{T} \): the larger \( \mathcal{T} \), the larger should the uniform deviation be. So, how can we quantify the size of \( \mathcal{T} \) for this problem? In the next section we will do precisely this – introduce a convenient, geometric measure of the sizes of sets in \( \mathbb{R}^n \), which is called Gaussian width.
1. Gaussian width

- **Definition.** Let $\mathcal{T} \subset \mathbb{R}^n$ be a bounded set, and $g$ be a standard normal random vector in $\mathbb{R}^n$, i.e. $g \sim \mathcal{N}(0, I_n)$. Then the quantities

\[
\omega(\mathcal{T}) := \mathbb{E} \sup_{x \in \mathcal{T}} \langle g, x \rangle \quad \text{and} \quad \gamma(\mathcal{T}) := \mathbb{E} \sup_{x \in \mathcal{T}} |\langle g, x \rangle|
\]

are called the **Gaussian width** of $\mathcal{T}$ and the **Gaussian complexity** of $\mathcal{T}$, respectively.

- Gaussian width and Gaussian complexity are closely related. Indeed, (Exercise)

\[
2\omega(\mathcal{T}) = \omega(\mathcal{T} - \mathcal{T}) = \mathbb{E} \sup_{x,y \in \mathcal{T}} \langle g, x - y \rangle = \mathbb{E} \sup_{x,y \in \mathcal{T}} |\langle g, x - y \rangle| = \gamma(\mathcal{T} - \mathcal{T}).
\]
Gaussian width has a natural geometric interpretation. Suppose $g$ is a unit vector in $\mathbb{R}^n$. Then a moment’s thought reveals that $\sup_{x,y \in T} \langle g, x - y \rangle$ is simply the width of $T$ in the direction of $g$, i.e. the distance between the two hyperplanes with normal $g$ that touch $T$ on both sides as shown in the figure. Then $2\omega(T)$ can be obtained by averaging the width of $T$ over all directions $g$ in $\mathbb{R}^n$. 

\[ \text{width} \]

\[ \langle g, x - y \rangle \]
• This reasoning is valid except where we assumed that $\mathbf{g}$ is a unit vector. Instead, for $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ we have $\mathbb{E}\|\mathbf{g}\|^2 = n$ and

$$\|\mathbf{g}\|_2 \approx \sqrt{n} \quad \text{with high probability.}$$

(Check both these claims using Bernstein’s inequality.) Thus, we need to scale by the factor $\sqrt{n}$. Ultimately, the geometric interpretation of the Gaussian width becomes the following: $\omega(\mathcal{T})$ is approximately $\sqrt{n}/2$ larger than the usual, geometric width of $\mathcal{T}$ averaged over all directions.

• A good exercise is to compute the Gaussian width and complexity for some simple sets, such as the unit balls of the $\ell_p$ norms in $\mathbb{R}^n$, which we denote by $\mathcal{B}_p^n = \{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq 1 \}$. We have

$$\gamma(\mathcal{B}_2^n) \sim \sqrt{n}, \quad \gamma(\mathcal{B}_1^n) \sim \sqrt{\log n}.$$  

For any finite set $\mathcal{T} \subset \mathcal{B}_2^n$, we have $\gamma(\mathcal{T}) \lesssim \sqrt{\log |\mathcal{T}|}$. The same holds for Gaussian width $\omega(\mathcal{T})$.  


2. Matrix deviation inequality

**Theorem 1 (Matrix deviation inequality)**

Let $A$ be an $m \times n$ matrix whose rows $A_{i,:}$ are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Let $\mathcal{T} \subset \mathbb{R}^n$ be a fixed bounded set. Then

$$
\mathbb{E} \sup_{x \in \mathcal{T}} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2 \gamma(\mathcal{T})
$$

where

$$
K = \max_i \|A_{i,:}\|_{\psi_2}
$$

is the maximal sub-gaussian norm of the rows of $A$.

### 2.1 Tail bound.

- It is often useful to have results that hold with high probability rather than in expectation. There exists a high-probability version of the matrix deviation inequality, and it states the following.
Let $u \geq 0$. Then the event

$$\sup_{x \in \mathcal{T}} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2[\gamma(\mathcal{T}) + u \cdot \text{rad}(\mathcal{T})]$$

holds with probability at least $1 - 2\exp(-u^2)$. Here $\text{rad}(\mathcal{T})$ is the radius of $\mathcal{T}$, defined as

$$\text{rad}(\mathcal{T}) := \sup_{x \in \mathcal{T}} \|x\|_2.$$ 

Since $\text{rad}(\mathcal{T}) \lesssim \gamma(\mathcal{T})$ we improve the bound as

$$\sup_{x \in \mathcal{T}} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \lesssim K^2u\gamma(\mathcal{T})$$

for all $u \geq 1$. This is a weaker but still a useful inequality. For example, we can use it to bound all higher moments of the deviation:

$$\left( \mathbb{E} \sup_{x \in \mathcal{T}} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right|^p \right)^{1/p} \leq C_pK^2\gamma(\mathcal{T})$$

where $C_p \leq C\sqrt{p}$ for $p \geq 1$. 
2.2 Deviation of squares

- It is sometimes helpful to bound the deviation of the square $\|Ax\|_2^2$ rather than $\|Ax\|_2$ itself.
- We can easily deduce the deviation of squares by using the identity
  
  $$a^2 - b^2 = (a - b)^2 + 2b(b - a)$$

  for $a = \|Ax\|_2$ and $b = \sqrt{m}\|x\|_2$.
- We have
  
  $$\mathbb{E} \sup_{x \in \mathcal{T}} \left| \|Ax\|_2^2 - m\|x\|_2^2 \right| \leq CK^4\gamma(\mathcal{T})^2 + CK^2\sqrt{m}\text{rad}(\mathcal{T})\gamma(\mathcal{T}).$$
2.3 Deriving Johnson-Lindenstrauss Lemma

- \( \mathcal{X} \subset \mathbb{R}^n \) and \( \mathcal{T} = \{ (x - y)/\|x - y\|_2 : x, y \in \mathcal{X} \} \). Then \( \mathcal{T} \) is finite and we have

\[
\gamma(\mathcal{T}) \lesssim \sqrt{\log |\mathcal{T}|} \leq \sqrt{\log |\mathcal{X}|^2} \lesssim \sqrt{\log |\mathcal{X}|}.
\]

Matrix deviation inequality and \( m \geq C\varepsilon^{-2} \log N \) then yield

\[
\sup_{x,y \in \mathcal{X}} \left| \frac{\| A(x - y) \|_2}{\|x - y\|_2} - \sqrt{m} \right| \lesssim \sqrt{\log N} \leq \varepsilon \sqrt{m}
\]

with high probability, say 0.99. Multiplying both sides by \( \|x - y\|_2/\sqrt{m} \), we can write the last bound as follows. With probability at least 0.99, we have

\[
(1 - \varepsilon)\|x - y\|_2 \leq \frac{1}{\sqrt{m}}\|Ax - Ay\|_2 \leq (1 + \varepsilon)\|x - y\|_2
\]

for all \( x, y \in \mathcal{X} \).
3. Covariance estimation

- We already showed that $N \sim n \log n$ samples are enough to estimate the covariance matrix of a general distribution in $\mathbb{R}^n$.
- We can do better if the distribution is sub-gaussian: we can get rid of the logarithmic oversampling and the boundedness condition.

**Theorem 2 (Covariance estimation for sub-gaussian distributions)**

Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$ with covariance matrix $\Sigma$. Suppose $\mathbf{X}$ is sub-gaussian, and more specifically for any $\mathbf{x} \in \mathbb{R}^n$

$$\|\langle \mathbf{X}, \mathbf{x} \rangle\|_{\psi_2} \lesssim \|\langle \mathbf{X}, \mathbf{x} \rangle\|_{L^2} = \|\Sigma^{1/2}\mathbf{x}\|_2.$$

Then, for every $N \geq 1$, we have

$$\mathbb{E}\|\Sigma_N - \Sigma\| \lesssim \|\Sigma\| \left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right).$$

- This result shows $N \sim \varepsilon^{-2}n$ gives $\mathbb{E}\|\Sigma_N - \Sigma\| \lesssim \varepsilon\|\Sigma\|$. 

Matrix deviation inequality   | DAMC Lecture 15   | May 29 - June 3, 2020   | 11 / 22
**Proof.** We first bring the random vectors $X, X_1, \ldots, X_N$ to the isotropic position. This can be done by a suitable linear transformation. You will easily check that there exists isotropic random vectors $Z, Z_1, \ldots, Z_N$ such that

$$X = \Sigma^{1/2} Z, \quad X_i = \Sigma^{1/2} Z_i, \quad i = 1, \ldots, N.$$  

The sub-gaussian assumption implies that

$$\|Z\|_{\psi_2} \lesssim 1.$$  

Then

$$\|\Sigma_N - \Sigma\| = \|\Sigma^{1/2} R_N \Sigma^{1/2}\| = \max_{\|x\|_2 = 1} \langle \Sigma^{1/2} R_N \Sigma^{1/2} x, x \rangle$$

where

$$R_N := \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^T - I_n.$$
Let $\mathcal{T} := \Sigma^{1/2}\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. We can rewrite $\|\Sigma_N - \Sigma\|$ as

$$
\|\Sigma_N - \Sigma\| = \max_{x \in \mathcal{T}} \langle R_N x, x \rangle = \max_{x \in \mathcal{T}} \frac{1}{N} \sum_{i=1}^N \langle Z_i, x \rangle^2 - \|x\|_2^2
= \frac{1}{N} \max_{x \in \mathcal{T}} \|A x\|_2^2 - N\|x\|_2^2.
$$

Now apply the matrix deviation inequality for squares to conclude that

$$
\mathbb{E} \|\Sigma_N - \Sigma\| \lesssim \frac{1}{N} \left( (\mathcal{T})^2 + \sqrt{N \operatorname{rad}(\mathcal{T}) \gamma(\mathcal{T})} \right).
$$

The radius and Gaussian width of the ellipsoid $\mathcal{T}$ are easy to compute:

$$
\operatorname{rad}(\mathcal{T}) = \|\Sigma\|^{1/2} \quad \text{and} \quad \gamma(\mathcal{T}) \leq (\operatorname{tr}(\Sigma))^{1/2}.
$$

By using $\operatorname{tr}(\Sigma) \leq n\|\Sigma\|$, we have

$$
\mathbb{E} \|\Sigma_N - \Sigma\| \lesssim \|\Sigma\| \left( \sqrt{\frac{n}{N}} + \frac{n}{N} \right). \quad \square
$$
3.1 Low-dimensional distributions

- We can show that much fewer samples are needed for covariance estimation of low-dimensional sub-gaussian distributions. Indeed, the proof actually yields

\[ \mathbb{E}\|\Sigma_N - \Sigma\| \lesssim \|\Sigma\| \left( \sqrt{\frac{r}{N}} + \frac{r}{N} \right) \]

where

\[ r = r(\Sigma^{1/2}) = \frac{\text{tr}\Sigma}{\|\Sigma\|} \]

is the stable rank of \( \Sigma^{1/2} \). This means that covariance estimation is possible with

\[ N \sim r \]

samples.
4. Underdetermined linear equations

- Suppose we need to solve a severely underdetermined system of linear equations: say, we have \( m \) equations in \( n \gg m \) variables

\[
Ax = y.
\]

- When the linear system is underdetermined, we can not find \( x \) with any accuracy, unless we know something extra about \( x \). So, let us assume that we do have some a-priori information. We can describe this situation mathematically by assuming that

\[
x \in \mathcal{K}
\]

where \( \mathcal{K} \subset \mathbb{R}^n \) is some known set in \( \mathbb{R}^n \) that describes anything that we know about \( x \) a-priori.

- Summarizing, here is the problem we are trying to solve. Determine a solution \( x = x(A, y, \mathcal{K}) \) to the underdetermined linear equation \( Ax = y \) as accurately as possible, assuming that \( x \in \mathcal{K} \).
4.1 An optimization approach

- We convert the set $\mathcal{K}$ into a function on $\mathbb{R}^n$ which is called the Minkowski functional of $\mathcal{K}$. This is basically a function whose level sets are multiples of $\mathcal{K}$.

- To define it formally, assume that $\mathcal{K}$ is star-shaped, which means that together with any point $x$, the set $\mathcal{K}$ must contain the entire interval that connects $x$ with the origin; see the figure for illustration.

![Diagram of star-shaped sets](image)

The set on the left (whose boundary is shown) is star-shaped, the set on the right is not.
The Minkowski functional of $\mathcal{K}$ is defined as

$$\|x\|_\mathcal{K} := \inf\{ t > 0 : x/t \in \mathcal{K} \}, \quad x \in \mathbb{R}^n.$$

If the set $\mathcal{K}$ is convex and symmetric about the origin, $\|x\|_\mathcal{K}$ is actually a norm on $\mathbb{R}^n$. (Exercise)

Now we propose the following way to solve the recovery problem: solve the optimization program

$$\min \|x'\|_\mathcal{K} \quad \text{subject to} \quad y = A x'.$$

Note that this is a very natural program: it looks at all solutions to the equation $y = A x'$ and tries to “shrink” the solution $x'$ toward $\mathcal{K}$. (This is what minimization of Minkowski functional is about.)

Also note that if $\mathcal{K}$ is convex, this is a convex optimization program, and thus can be solved effectively by one of the many available numeric algorithms.
• The main question we should now be asking is – would the solution to this program approximate the original vector \( \mathbf{x} \)? The following result bounds the approximation error for a probabilistic model of linear equations.

**Theorem 3 (Recovery by optimization)**

Assume that \( \mathbf{A} \) is an \( m \times n \) random matrix whose rows \( \mathbf{A}_{i,:} \) are independent, isotropic and sub-gaussian random vectors in \( \mathbb{R}^n \). The solution \( \hat{\mathbf{x}} \) of the optimization problem satisfies

\[
\mathbb{E} \| \hat{\mathbf{x}} - \mathbf{x} \|_2 \lesssim \frac{\omega(K)}{\sqrt{m}},
\]

where \( \omega(K) \) is the Gaussian width of \( K \).

**Proof.** Both the original vector \( \mathbf{x} \) and the solution \( \hat{\mathbf{x}} \) are feasible vectors for the optimization program. Then \( \| \hat{\mathbf{x}} \|_K \leq \| \mathbf{x} \|_K \leq 1 \). Thus both \( \hat{\mathbf{x}}, \mathbf{x} \in K \).
By $A\hat{x} = Ax = y$, we have $A(\hat{x} - x) = 0$. Let us apply matrix deviation inequality for $\mathcal{T} := \mathcal{K} - \mathcal{K}$. It gives

$$\sup_{u,v \in \mathcal{K}} \left| \| A(u - v) \|_2 - \sqrt{m} \| u - v \|_2 \right| \lesssim \gamma(\mathcal{T}) = 2\omega(\mathcal{K}).$$

Substitute $u = \hat{x}$ and $v = x$ here. We may do this since, as we noted above, both these vectors belong to $\mathcal{K}$. But then the term $\| A(u - v) \|_2$ will equal zero. It disappears from the bound, and we get

$$\mathbb{E} \sqrt{m} \| \hat{x} - x \|_2 \lesssim \omega(\mathcal{K}).$$

Dividing both sides by $\sqrt{m}$ we complete the proof.

This theorem says that a signal $x \in \mathcal{K}$ can be efficiently recovered from

$$m \sim \omega(\mathcal{K})^2$$

random linear measurements.
5. Sparse recovery

- Suppose we know that the signal $\mathbf{x}$ is sparse, which means that only a few coordinates of $\mathbf{x}$ are nonzero. As before, our task is to recover $\mathbf{x}$ from the random linear measurements given by the vector

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

where $\mathbf{A}$ is an $m \times n$ random matrix.

- The number of nonzero coefficients of a vector $\mathbf{x} \in \mathbb{R}^n$, or the sparsity of $\mathbf{x}$, is often denoted $\|\mathbf{x}\|_0$. This is reminiscent of the notation for the $\ell_p$ norm $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, and for a reason. You can quickly check that

$$\|\mathbf{x}\|_0 = \lim_{p \to 0} \|\mathbf{x}\|_p.$$

- Keep in mind that neither $\|\mathbf{x}\|_0$ nor $\|\mathbf{x}\|_p$ for $0 < p < 1$ are actually norms on $\mathbb{R}^n$, since they fail triangle inequality.
Our first attempt to recover \( \mathbf{x} \) is to try the following optimization problem:

\[
\min \| \mathbf{x}' \|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}'.
\]

This is sensible because this program selects the sparsest feasible solution. But there is an implementation caveat: the function \( f(\mathbf{x}) = \| \mathbf{x} \|_0 \) is highly non-convex and even discontinuous. There is simply no known algorithm to solve the optimization problem efficiently.

To overcome this difficulty, we use

\[
\min \| \mathbf{x}' \|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}'.
\]

This is a convexification of the non-convex program, and a variety of numeric convex optimization methods are available to solve it efficiently.

We will now show that an \( s \)-sparse signal \( \mathbf{x} \in \mathbb{R}^n \) can be efficiently recovered from \( m \sim s \log n \) random linear measurements.
Theorem 4 (Sparse recovery by optimization)

Assume $A$ is a random matrix as in Theorem 3. If an unknown vector $x \in \mathbb{R}^n$ has at most $s$ non-zero coordinates, i.e. $\|x\|_0 \leq s$, then the solution $\hat{x}$ of the $\ell_1$ optimization program satisfies

$$\mathbb{E}\|\hat{x} - x\|_2 \lesssim \sqrt{(s \log n)/m}\|x\|_2.$$  

Proof. Cauchy-Schwarz inequality shows that $\|x\|_1 \leq \sqrt{s}\|x\|_2$. Denote the unit ball of the $\ell_1$ norm in $\mathbb{R}^n$ by $B_1^n$. Then we can rewrite $\|x\|_1 \leq \sqrt{s}\|x\|_2$ as the inclusion

$$x \in \sqrt{s}\|x\|_2 \cdot B_1^n := K.$$  

By the Gaussian width of $B_1^n$, we have

$$\omega(K) = \sqrt{s}\|x\|_2 \cdot \omega(B_1^n) \leq \sqrt{s}\|x\|_2 \cdot \gamma(B_1^n) \leq \sqrt{s}\|x\|_2 \cdot \sqrt{\log n}.$$  

Substitute this in Theorem 3 and complete the proof.