

Lecture 14: Matrix deviation inequality



School of Mathematical Sciences, Xiamen University

1. Gaussian width and Gaussian complexity

- Assume that \mathbf{A} is an $m \times n$ random matrix whose rows are independent, mean zero, isotropic and sub-gaussian random vectors in \mathbb{R}^n . For example,

$$\mathbf{A} = \text{randn}(m, n).$$

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

$$\begin{aligned}\mathbb{E}\|\mathbf{A}\mathbf{x}\|_2^2 &= \mathbb{E} \sum_{i=1}^m (\mathbf{A}_{i,:}\mathbf{x})^2 = \sum_{i=1}^m \mathbb{E}(\mathbf{A}_{i,:}\mathbf{x})^2 \\ &= \sum_{i=1}^m \mathbf{x}^\top \mathbb{E}(\mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}) \mathbf{x} \\ &= m\|\mathbf{x}\|_2^2.\end{aligned}$$

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

$$\|\mathbf{Ax}\|_2 \approx \sqrt{m}\|\mathbf{x}\|_2$$

with high probability. (Recall Johnson–Lindenstrauss Lemma)

- A natural problem is, for all \mathbf{x} in some fixed set $\mathcal{T} \subset \mathbb{R}^n$, how large is the average uniform deviation:

$$\mathbb{E} \sup_{\mathbf{x} \in \mathcal{T}} \left| \|\mathbf{Ax}\|_2 - \sqrt{m}\|\mathbf{x}\|_2 \right|?$$

- 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? Gaussian width: a geometric measure of the sizes of sets in \mathbb{R}^n .

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

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

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

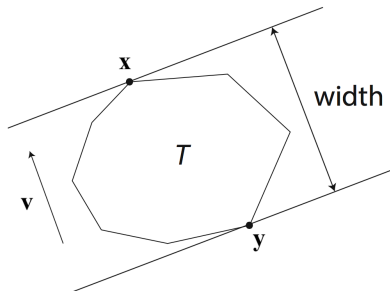
- Gaussian width and Gaussian complexity are closely related. Indeed, we have (see Vershynin's HDP book)

$$\begin{aligned} \omega(\mathcal{T} - \mathcal{T}) &= \mathbb{E} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{T}} \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle = 2\omega(\mathcal{T}) \\ &= \mathbb{E} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{T}} |\langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle| \\ &= \gamma(\mathcal{T} - \mathcal{T}), \end{aligned}$$

$\omega(\mathcal{T}) \leq \gamma(\mathcal{T})$, and if \mathcal{T} contains the origin, then $\gamma(\mathcal{T}) \leq 2\omega(\mathcal{T})$.

- Geometric interpretation.

Suppose \mathbf{v} is a unit vector in \mathbb{R}^n . Then $\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{T}} \langle \mathbf{v}, \mathbf{x} - \mathbf{y} \rangle$ is simply the width of \mathcal{T} in the direction of \mathbf{v} , i.e. the distance between the two hyperplanes with normal \mathbf{v} that touch \mathcal{T} on both sides as shown in the figure.



For $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, we have $\mathbb{E}\|\mathbf{g}\|_2^2 = n$ and $\|\mathbf{g}\|_2 \approx \sqrt{n}$ with high probability. Then, $\omega(\mathcal{T})$ is approximately $\sqrt{n}/2$ larger than the usual, geometric width of \mathcal{T} averaged over all directions.

- Gaussian complexity for the unit balls $\mathcal{B}_p^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq 1\}$:

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

For any finite set $\mathcal{T} \subset \mathcal{B}_2^n$, we have $\gamma(\mathcal{T}) \lesssim \sqrt{\log |\mathcal{T}|}$.

2. Matrix deviation inequality

Theorem 1 (Matrix deviation inequality)

Let \mathbf{A} be an $m \times n$ matrix whose rows $\mathbf{A}_{i,:}$ 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

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

where

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

is the maximal sub-gaussian norm of the rows of \mathbf{A} .

2.1 Deriving Johnson–Lindenstrauss Lemma

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

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

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

$$(1 - \varepsilon)\|\mathbf{x} - \mathbf{y}\|_2 \leq \frac{1}{\sqrt{m}} \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2 \leq (1 + \varepsilon)\|\mathbf{x} - \mathbf{y}\|_2$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

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 Σ . Suppose \mathbf{X} is sub-gaussian, mean zero, 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 $\mathbb{E} \|\Sigma_N - \Sigma\| \lesssim \varepsilon \|\Sigma\|$.

3.1 Low-dimensional distributions

- We can show that much fewer samples are needed for covariance estimation of low-dimensional sub-gaussian distributions. We have

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

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 systems

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

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}, \quad \mathbf{b} \in \mathbb{R}^m.$$

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

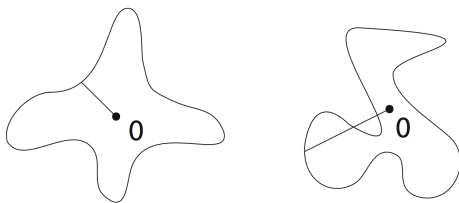
$$\mathbf{x} \in \mathcal{K}$$

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

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

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 \mathbf{x} , the set \mathcal{K} must contain the entire interval that connects \mathbf{x} with the origin; see the figure for illustration.



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

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

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

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

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

It looks at all solutions to the equation $\mathbf{b} = \mathbf{A}\mathbf{x}$ and tries to “shrink” the solution \mathbf{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 problem, 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 problem 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, mean zero, isotropic and sub-gaussian random vectors in \mathbb{R}^n . The solution $\hat{\mathbf{x}}$ of the optimization problem satisfies

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

where $\omega(\mathcal{K})$ is the Gaussian width of \mathcal{K} .

Proof. By $\mathbf{x} \in \mathcal{K}$ and the optimality, we have $\|\hat{\mathbf{x}}\|_{\mathcal{K}} \leq \|\mathbf{x}\|_{\mathcal{K}} \leq 1$, which implies $\hat{\mathbf{x}} \in \mathcal{K}$. By $\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{x} = \mathbf{b}$, we have $\mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}) = \mathbf{0}$.

Let us apply matrix deviation inequality for $\mathcal{T} := \mathcal{K} - \mathcal{K}$. It gives

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

Substitute $\mathbf{u} = \hat{\mathbf{x}}$ and $\mathbf{v} = \mathbf{x}$ here. By $\mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}) = \mathbf{0}$, we get

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

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

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

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

random linear measurements.

4.2 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

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

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

- The number of nonzero entries of a vector $\mathbf{x} \in \mathbb{R}^n$, or the sparsity of \mathbf{x} , is often denoted $\|\mathbf{x}\|_0$. Recall that $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. You can quickly check that

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

- 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.

- The problem is:

$$\min \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{b} = \mathbf{A}\mathbf{x},$$

which selects the sparsest feasible solution. The function $\|\mathbf{x}\|_0$ is highly non-convex and even discontinuous. No known algorithm exists to solve this problem efficiently.

- To overcome this difficulty, we use

$$\min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{b} = \mathbf{A}\mathbf{x}.$$

This is a convexification of the non-convex problem, 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 \mathbf{A} is a random matrix as in Theorem 3. If an unknown vector $\mathbf{x} \in \mathbb{R}^n$ has at most s non-zero coordinates, i.e. $\|\mathbf{x}\|_0 \leq s$, then the solution $\hat{\mathbf{x}}$ of the ℓ_1 optimization problem satisfies

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

Proof. Cauchy–Schwarz inequality shows that $\|\mathbf{x}\|_1 \leq \sqrt{s} \|\mathbf{x}\|_2$. Denote the unit ball of the ℓ_1 norm in \mathbb{R}^n by \mathcal{B}_1^n . Then we can rewrite $\|\mathbf{x}\|_1 \leq \sqrt{s} \|\mathbf{x}\|_2$ as the inclusion

$$\mathbf{x} \in \sqrt{s} \|\mathbf{x}\|_2 \cdot \mathcal{B}_1^n := \mathcal{K}.$$

By the Gaussian complexity $\gamma(\mathcal{B}_1^n) \lesssim \sqrt{\log n}$, we have

$$\omega(\mathcal{K}) = \sqrt{s} \|\mathbf{x}\|_2 \cdot \omega(\mathcal{B}_1^n) \leq \sqrt{s} \|\mathbf{x}\|_2 \cdot \gamma(\mathcal{B}_1^n) \lesssim \sqrt{s} \|\mathbf{x}\|_2 \cdot \sqrt{\log n}.$$

Substitute this in Theorem 3 and complete the proof. □