1. Span

$$V_8 = \sum_{i=1}^{7} \alpha_i V_i \Rightarrow V_8 \in \text{Span}\{V_1, \cdots, V_7\}$$

$$V_8 - \sum_{i=1}^{7} \alpha_i V_i = 0$$ 若有解，则也span.

2. Inner Product (Dot product)

$$\langle u, v \rangle = u \cdot v = \sum_{i=1}^{n} u_i \cdot v_i$$

$$\|u\|_2 = \sqrt{\sum_{i=1}^{n} u^2} = \sqrt{u \cdot u}$$ = the vector's length

$$|u \cdot v| \leq \|u\|_2 \cdot \|v\|_2$$

3. Othogonal & Orthonormal

两个向量 $\mathbf{u}$ 和 $\mathbf{v}$ 被称为正交，当且仅当它们的内积为零：

$$\mathbf{u} \cdot \mathbf{v} = 0$$

几何解释：

1. 正交意味着这两个向量在几何上是垂直的（夹角 $\theta = 90^\circ$）。
2. 这是因为内积公式：

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos \theta$$

当 $\cos\theta = 0$ 时，$\theta = 90^\circ$，即向量相互垂直。

标准正交（Orthonormal）

如果两个向量不仅正交（即 $\mathbf{u} \cdot \mathbf{v} = 0$），而且它们的长度（范数）为 1，则称它们为**标准正交**：

$$\|\mathbf{u}\| = \|\mathbf{v}\| = 1$$

几何解释：

1. 标准正交表示两个向量不仅彼此垂直，而且是单位向量（长度为 1）。

4. Basis

basis is the minimal representation of subspace (all vectors are linearly independent and every other vector in the subspace can represented as a linear combination of the basis vectors). A basis is orthonormal if and only if all the vectors in the basis are orthogonal to each other and have a length of 1.

A basis $B = \{v_1, v_2, \ldots, v_m\}$ for a subspace $V \subset \mathbb{R}^n$ is called **orthonormal** if:

$$\langle v_i, v_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

We can define a vector $u \in V$, where V has an orthonormal basis of $\{v_1, v_2, ..., v_m\}$. Since we know u is a linear combination of the vectors in the basis, we can write u as

$$u = \sum_{i=1}^{m} \alpha_i v_i \quad \text{where} \quad \alpha_i \in \mathbb{R}$$

To find these α values, we can simply take the dot product of u with each vector in the orthonormal basis v_i.

$$\begin{aligned} \langle u, v_j \rangle &= \langle \sum_{i=1}^{m} \alpha_i v_i, v_j \rangle \\ &= \sum_{i=1}^{m} \alpha_i \langle v_i, v_j \rangle \\ &= \alpha_j \end{aligned}$$

We can also define this dot project of u and v_j as the length of the projection of u onto v_j.

We can then continue to further prove that the 2-norm (length) of a vector is simply the sum of these α_i values squared. Specifically, that

$$\sum_{i=1}^{m} \alpha_i^2 = \|u\|_2^2$$

5. Projection

10.1.1 Projection Formula

The projection of a vector $u \in \mathbb{R}^n$ onto a vector v_j is given by:

$$\text{proj}(u, v_j) = \frac{\langle u, v_j \rangle}{\langle v_j, v_j \rangle} v_j$$

If W is the subspace spanned by orthonormal vectors $v_1, \ldots, v_k$, then the projection of u onto W is:

$$\text{proj}(u, W) = \sum_{i=1}^{k} \langle u, v_i \rangle v_i$$

Consider a vector $v = (2, 1, 2)$ and the xy-plane spanned by $(1, 0, 0)$ and $(0, 1, 0)$. The projection of v onto the xy-plane is:

$$\text{proj}(v, \text{xy-plane}) = \langle v, (1,0,0) \rangle \cdot (1,0,0) + \langle v, (0,1,0) \rangle \cdot (0,1,0) = (2,1,0)$$

6. Gram-Schmidt Algorithm

Given vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ in $\mathbb{R}^n$ that form a basis for a subspace V, the Gram-Schmidt process outputs orthonormal vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ that span the same subspace V as follows:

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$$

$$\mathbf{u}_2 = \frac{\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1}{\|\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1\|}$$

$$\mathbf{u}_k = \frac{\mathbf{v}_k - \sum_{i=1}^{k-1} \langle \mathbf{v}_k, \mathbf{u}_i \rangle \mathbf{u}_i}{\|\mathbf{v}_k - \sum_{i=1}^{k-1} \langle \mathbf{v}_k, \mathbf{u}_i \rangle \mathbf{u}_i\|}$$

Error: $\sum_{i=1}^{n} (v_i - w_i)^2$

Given vectors $v_1, \ldots, v_8$ in $\mathbb{R}^n$, we want to address the following questions:

- **Q**: Is v_8 in the span of $v_1, \ldots, v_7$?

We want to minimize:

$$\min \|v_8 - w\| \quad \text{where} \quad w \in \text{span}(v_1, \ldots, v_7)$$

To solve this, we will use **Gram-Schmidt Orthogonalization** to construct an orthonormal basis $u_1, \ldots, u_7$. The projection of v_8 onto the subspace spanned by $u_1, \ldots, u_7$ will give us the vector w that minimizes the distance, with our notion of "distance" defined as the 2-norm.

通过正交基的性质，投影可以直接用公式计算：

$$w = \sum_{i=1}^{7} \langle v_8, u_i \rangle u_i \longrightarrow w = \sum_{i=1}^{k} \beta_i w_i, \quad \beta_i = \langle v_8, w_i \rangle.$$

7. Eigenvalue & Eigenvector

Suppose $A \in \mathbb{R}^{n \times n}$, then $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ is said to be an eigenvector-eigenvalue pair if:

$$Ax = \lambda x$$

The eigenvector v corresponding to an eigenvalue must be non-zero by definition.

Find eigenvalues:

$M_\lambda = A - \lambda I$

$\det(M_\lambda) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i V_i = 0$

这也意味着 M_λ is noninvertible

8. Invertible

一个**矩阵** A **是可逆矩阵**（invertible matrix 或 nonsingular matrix），如果存在另一个矩阵 A^{-1}（称为 A 的逆矩阵），满足以下两个条件：

$$A \cdot A^{-1} = I \quad 和 \quad A^{-1} \cdot A = I$$

其中，I 是单位矩阵（identity matrix），其对角线元素全为 1，其他元素全为 0。

1. **矩阵可逆的核心思想**：
   一个矩阵 A 可逆，意味着我们可以"还原"它的线性变换——对于任何向量 v，如果知道 Av，就可以用 A^{-1} 找回 v。
   例如：如果 $Av = b$，我们可以通过 $v = A^{-1}b$ 找回 v。
2. **可逆矩阵是全秩矩阵**：
   如果 A 是 $n \times n$ 方阵，它的行和列必须是线性无关的（即满秩，rank = n），才能可逆。
3. **行列式的作用**：
   一个 $n \times n$ 方阵 A 的**行列式**（determinant）可以判断它是否可逆：
   - 如果 $\det(A) \neq 0$，则 A 可逆。
   - 如果 $\det(A) = 0$，则 A 不可逆（也称为奇异矩阵，singular matrix）。

9. Special theorem for symmetric matrix

If $A \in \mathbb{R}^{n \times n}$ is symmetric, then all eigenvalues are real.

Further, A can be expressed as:

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T \quad \text{(Spectral decomposition)}$$

where $v_1, \ldots, v_n$ are orthonormal unit vectors, and:

$$Av_i = \lambda_i v_i \quad \forall\, 1 \leq i \leq n$$

$\langle v_i, v_j \rangle = \begin{cases} = 1 & i = j \\ = 0 & i \neq j \end{cases}$

ex: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 7 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} [1\ 0\ 0] - 5 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} [0\ 1\ 0] \cdots$

$$A^2 = \sum_{i,j=1}^{n} \lambda_i \lambda_j \left(v_i (v_i^T v_j) v_j^T \right) = \sum_{i=1}^{n} \lambda_i^2 v_i v_i^T$$

10. Rayleigh Quotient

Rayleigh 商是一个实数值函数，用于描述向量和对称矩阵 A 之间的关系。对于任意非零向量 $x \in \mathbb{R}^n$，其定义为：

$$R(x) = \frac{x^T A x}{x^T x}$$

其中：

- A 是对称矩阵；
- $x^T A x$ 是二次型；
- $x^T x = \|x\|^2$ 是向量 x 的平方范数。

性质

- **缩放不变性**：Rayleigh 商对 x 的缩放不敏感。如果 $y = cx$（$c \neq 0$ 是标量），则：

$$R(y) = \frac{y^T A y}{y^T y} = \frac{(cx)^T A (cx)}{(cx)^T (cx)} = \frac{c^2 x^T A x}{c^2 x^T x} = R(x)$$

因此，Rayleigh 商只依赖于 x 的方向，而与其大小无关。

假设 $\|x\|_2 = x^T x = \sum_{i=1}^{n} \alpha_i^2 = 1$

则 $R = x^T A x = \left(\sum_{i=1}^{n} \alpha_i v_i\right)^T A \left(\sum_{j=1}^{n} \alpha_j v_j\right)$

$= \sum_{i=1}^{n} \lambda_i \alpha_i^2$

注：x 可以被写成 $\sum_{i=1}^{n} \alpha_i V_i$

5. 优化问题

目标：找到一个单位向量 x 使得 $x^T A x$ 最大化。

利用 Rayleigh 商公式：

$$R(x) = \sum_{i=1}^{n} \lambda_i \alpha_i^2$$

- 由于 $\alpha_i^2 \geq 0$ 且 $\sum_{i=1}^{n} \alpha_i^2 = 1$，$R(x)$ 的值由最大的特征值 $\lambda_{\max}$ 主导。
- 为了最大化 $R(x)$，需要让 $\alpha_1^2 = 1$，即 x 完全沿最大特征值 $\lambda_{\max}$ 的特征向量方向 v_1 排布。

解：最优 x 是 v_1（最大特征值的特征向量）。

The expression $\sum_{i=1}^{n} \alpha_i^2 \lambda_i$ is maximized when $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = \cdots = \alpha_n = 0$. This implies that we are allotting our entire budget (recall that $\sum_{i=1}^{n} \alpha_i = 1$) to our largest eigenvalue λ_1, and in doing so, we make $x = v_1$.

11. Variational Characteristic of Eigenvalues

Now, suppose we add an additional constraint to our optimization problem.

$$\begin{aligned} \max. \quad & x^T A x \\ \text{s.t.} \quad & x^T x = 1 \\ & < x, v_1 >= 0 \end{aligned}$$

3. 递归优化与一般性结果

通过类似的步骤，可以递归地限制 x 正交于之前的特征向量 $v_1, v_2, \ldots, v_{k-1}$，从而找到下一个最大的特征值 λ_k 和对应特征向量 v_k。总结如下：

- 最大特征值：$\lambda_1 = \max_x x^T A x, \quad \text{where } x = v_1.$
- 第二大特征值：$\lambda_2 = \max_{x \perp v_1} x^T A x, \quad \text{where } x = v_2.$
- 一般情形：$\lambda_k = \max_{x \perp \{v_1, v_2, \ldots, v_{k-1}\}} x^T A x, \quad \text{where } x = v_k.$

12. Singular Value Decomposition (SVD)

1. 什么是奇异值分解？

假设我们有一个矩阵 $A \in \mathbb{R}^{m \times n}$，其中 $n \geq m$。奇异值分解 (SVD) 将矩阵 A 分解为以下形式：

$$A = \sum_{i=1}^{m} \sigma_i u_i v_i^T,$$

where each term $\sigma_i u_i v_i^T$ is a rank-1 matrix, $u_i \in \mathbb{R}^m$, and $v_i \in \mathbb{R}^n$. Furthermore, the vectors $u_1, \ldots, u_m$ are orthonormal, as are $v_1, \ldots, v_m$, and $\sigma_i \geq 0$.

$$\begin{aligned} AA^T &= \left(\sum_{i=1}^{m} \sigma_i u_i v_i^T\right)\left(\sum_{j=1}^{m} \sigma_j v_j u_j^T\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i \sigma_j u_i v_i^T v_j u_j^T \\ &= \sum_{i=1}^{m} \sigma_i^2 u_i u_i^T \qquad \text{since } v_i^T v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$A^T A = \sum_{i=1}^{m} \sigma_i^2 v_i v_i^T$

$v_1 \cdots v_m$ are eigenvectors; $\sigma_1^2 \cdots \sigma_m^2$ are eigenvalues.

Step 1: Singular Value Decomposition (SVD) and the Representation of A

1. **Definition of SVD:** A matrix $A \in \mathbb{R}^{m \times n}$ can be written as:

$$A = U \Sigma V^T$$

where:

- $U = [u_1, u_2, \ldots, u_m] \in \mathbb{R}^{m \times m}$ contains orthonormal left singular vectors,
- $V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times n}$ contains orthonormal right singular vectors,
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with singular values $\sigma_1, \sigma_2, \ldots$ on the diagonal.

2. **SVD in Summation Form:** Using the above, A can be expressed as:

$$A = \sum_i \sigma_i u_i v_i^T$$

So:

$$\|A\|_F^2 = \sum_{j,k}\left(\sum_i \sigma_i (u_i)_j (v_i)_k\right)^2$$

(b) Expanding the Square

Expand the square:

$$\|A\|_F^2 = \sum_{j,k}\sum_{i,\ell} \sigma_i \sigma_\ell (u_i)_j (v_i)_k (u_\ell)_j (v_\ell)_k$$

(c) Orthogonality of u_i and v_i

- The vectors u_i are orthonormal, so:

$$\sum_j (u_i)_j (u_\ell)_j = \delta_{i\ell}$$

where $\delta_{i\ell}$ is the Kronecker delta (1 if $i = \ell$, 0 otherwise).

- Similarly, the vectors v_i are orthonormal, so:

$$\sum_k (v_i)_k (v_\ell)_k = \delta_{i\ell}$$

Substitute these orthogonality conditions back into the expression:

$$\|A\|_F^2 = \sum_{j,k}\sum_{i,\ell} \sigma_i \sigma_\ell \delta_{i\ell} \delta_{i\ell} = \sum_i \sigma_i^2$$

Thus:

$$\|A\|_F^2 = \sum_i \sigma_i^2$$

(arranged in decreasing order of σ_i's), where:

- $\sigma_1, \ldots, \sigma_m$ are the singular values, which are non-negative real numbers ($\sigma_i \geq 0$).
- $u_1, \ldots, u_m$ are orthonormal unit vectors in $\mathbb{R}^m$.
- $v_1, \ldots, v_n$ are orthonormal unit vectors in $\mathbb{R}^n$, forming the right singular vectors.

13. Direction of Maximum Variance

我们希望找到一个单位向量 x（即 $\|x\|_2 = 1$），使得 $\|Ax\|_2$ 最大化。

这个问题可以写成：

$$\max_{\|x\|_2=1} \|Ax\|_2,$$

或者等价地写成：

$$\max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

而 $\|Ax\|_2^2 = \langle Ax, Ax \rangle$ 和 $\|x\|_2^2 = \langle x, x \rangle$，因此问题变为：

$$\max_{x \neq 0} \sqrt{\frac{\langle Ax, Ax \rangle}{\langle x, x \rangle}}.$$

简化问题

进一步简化，可以写成：

$$\max_{x \neq 0} \sqrt{\frac{(Ax)^T (Ax)}{x^T x}}.$$

再简化为：

$$\max_{x \neq 0} \sqrt{\frac{x^T (A^T A) x}{x^T x}}.$$

到这里，我们可以将问题归约为：

$$\max_{x \neq 0} \frac{x^T (A^T A) x}{x^T x}.$$

这个形式与**Rayleigh 商**相同，最大值等于 $A^T A$ 的最大特征值。

特征值与奇异值的关系

在 SVD 中，$A^T A$ 的特征值是奇异值的平方：

$$\lambda_1 = \sigma_1^2, \quad \lambda_2 = \sigma_2^2, \ldots.$$

因此，$A^T A$ 的最大特征值是 σ_1^2，对应的特征向量是 v_1。于是：

$$\max_{x \neq 0} \frac{x^T (A^T A) x}{x^T x} = \sigma_1^2.$$

方向与最大奇异值

为了达到这个最大值，单位向量 x 必须是 $A^T A$ 的主特征向量 v_1。因此：

$$\max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1, \quad \text{where } x = v_1.$$

---

(1) Rayleigh 商定理 Rayleigh 补充

Rayleigh 商定理指出：

1. 对于对称矩阵 $A^T A$，Rayleigh 商 $R(x) = \frac{x^T (A^T A) x}{x^T x}$ 的最大值等于 $A^T A$ 的最大特征值。
2. 最大值对应的 x 是矩阵 $A^T A$ 的主特征向量（即对应最大特征值的特征向量）。

因此：

$$\max_{x \neq 0} \frac{x^T (A^T A) x}{x^T x} = \lambda_1,$$

其中 λ_1 是 $A^T A$ 的最大特征值。

---

(2) 奇异值与特征值的关系

在奇异值分解（SVD）中：

- $A^T A$ 的特征值是奇异值的平方：

$$\lambda_i = \sigma_i^2.$$

- σ_1 是矩阵 A 的最大奇异值，因此 $\lambda_1 = \sigma_1^2$。

Similarly to Sec. 12.3, we can find the next largest singular value by maximizing $\|Ax\|_2$ when x is orthogonal to the previous singular vectors. Define σ_1 as:

$$\sigma_1 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

Now suppose $\langle x, v_1 \rangle = 0$ and let $x = \sum_{j=2}^{n} \alpha_j v_j$. Then,

$$\begin{aligned} Ax &= \left(\sum_{i=1}^{n} \sigma_i u_i v_i^T\right)\left(\sum_{j=2}^{n} \alpha_j v_j\right) \\ &= \sum_{j=2}^{m} \alpha_j \sigma_j u_j. \end{aligned}$$

To maximize $\|Ax\|_2$, set $\alpha_2 = 1$. Then the maximum value becomes σ_2 when $x = v_2$, and so on...

Therefore,

$$\beta_1 = 1, \beta_2 = \beta_3 = \cdots = \beta_m = 0$$
$$\|A\mathbf{x}\|_2 = \sigma_1$$

Hence,

$$\sigma_1 = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2 \quad ; \quad v_1 = \arg\max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2$$

Similarly,

$$\sigma_2 = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1, \langle \mathbf{x}, v_1 \rangle = 0} \|A\mathbf{x}\|_2 \quad ; \quad v_2 = \arg\max_{\mathbf{x}:\|\mathbf{x}\|_2=1, \langle \mathbf{x}, v_1 \rangle = 0} \|A\mathbf{x}\|_2$$
$$\sigma_k = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1, \langle \mathbf{x}, v_1 \rangle = \cdots = \langle \mathbf{x}, v_{k-1} \rangle = 0} \|A\mathbf{x}\|_2 \quad ; \quad v_k = \arg\max_{\mathbf{x}:\|\mathbf{x}\|_2=1, \langle \mathbf{x}, v_1 \rangle = \cdots = \langle \mathbf{x}, v_{k-1} \rangle = 0} \|A\mathbf{x}\|_2$$

14. SVD的奇怪定理

Given matrix A, the matrix A_k of rank-k which minimizes both the 2-norm and Frobenius norm is given by

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

where

$$A = \sum_{i=1}^{m} \sigma_i u_i v_i^T$$

给定矩阵 A 的奇异值分解 (Singular Value Decomposition, SVD)，存在一个秩为 k 的矩阵 A_k，它是以下两个范数意义上的最佳近似：

1. **2-范数 (Spectral Norm)**：最大奇异值对应的近似误差最小；
2. **Frobenius 范数**：所有元素的平方和的近似误差最小。

15. Recovering Partitions

假设有一组顶点 V，被分成两个隐藏的部分 S 和 $\overline{S}$，满足：

- $|S| = |\overline{S}| = \frac{N}{2}$，即两部分大小相等。

边生成规则：

- 如果 $i, j \in S$ 或 $i, j \in \overline{S}$（即两点在同一部分中），则以概率 p 生成一条边；
- 如果 $i \in S, j \in \overline{S}$ 或 $i \in \overline{S}, j \in S$（即两点在不同部分中），则以概率 q 生成一条边。

其中 $p > q$。

当 $p - q \geq \frac{100}{\sqrt{n}}$ 时，特征值分解可以成功恢复几乎所有的划分（即错误率控制在 0.01 的水平）。

Chernoff Bound:
If $x_1, x_2, \ldots, x_n$ are random variables in [-1, 1], then $Pr[(\sum_i x_i - \mathbf{E}[\sum_i x_i]) \geq t] \leq e^{-\frac{t^2}{4n}}$

We aim to recover the partition $(S, \bar{S})$ when:

$$p \geq q + \frac{\log 5n}{\sqrt{n}}.$$

Fact 1: The all-ones matrix $J_{n/2}$ of size $n/2 \times n/2$ has exactly one non-zero eigenvalue $\lambda = n/2$, with corresponding eigenvector $\mathbf{v} = \frac{1}{\sqrt{n/2}}\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. All other eigenvalues are zero.

Fact 2: If a matrix B has eigenvector v with eigenvalue λ, then for any scalar c, the matrix $B - cI$ has eigenvector v with eigenvalue $\lambda - c$.

If A, B belong to the same cluster, then:

$$\varepsilon_c = \begin{cases} 1 & \text{if both } (A,C), (B,C) \text{ are present/absent} \\ 0 & \text{otherwise} \end{cases}$$

$$E\left[\sum_c \varepsilon_c\right] \geq 0.81 \cdot (n-2)$$

$$\Pr\left(\sum_c \varepsilon_c \leq \frac{n-2}{2}\right) \leq \Pr\left[\left|\sum_c \varepsilon_c - E\left(\sum_c \varepsilon_c\right)\right| \geq 0.31 \cdot (n-2)\right]$$

$$\leq e^{-\frac{(0.31)^2 \cdot (n-2)^2}{4 \cdot (n-2)}} \approx e^{-0.02n}$$

$$= \Pr\left(\sum_c \varepsilon_c \geq \frac{n-2}{2}\right) \geq 1 - e^{-0.02n}$$

If A, B belong to different clusters:

$$E\left[\sum_c \varepsilon_c\right] \leq 0.18 \cdot (n-2)$$

$$\Rightarrow \quad \Pr\left(\sum_c \varepsilon_c \geq \frac{(n-2)}{2}\right) \leq e^{-0.02 \cdot n}$$

As long as:

$$\frac{n(p-q)}{2} \gg \sqrt{n(p+q)},$$

which simplifies to:

$$\frac{p-q}{2} \gg \frac{1}{\sqrt{n}},$$

we can successfully recover the communities using the spectral method described above.

16. Random Matrix Analysis

1. **随机矩阵 Δ 的定义**：

   1 1 p q

   $$\Delta = C - C' = M - A.$$

   矩阵 Δ 捕捉了随机噪声的特性。

   - 每个元素 Δ_{ij} 是随机的，但独立（对称性除外）。
   - 期望 $\mathbb{E}[\Delta_{ij}] = 0$。
   - 方差 $\mathrm{Var}(\Delta_{ij})$：

   $$\mathrm{Var}(\Delta_{ij}) = \begin{cases} p(1-p), & \text{如果 } i \neq j \text{ 且在同一社区}, \\ q(1-q), & \text{如果 } i \neq j \text{ 且在不同社区}. \end{cases}$$

2. **矩阵分析工具**：

   - **Tao-Vu 定理**：
     如果矩阵 Δ 是对称的随机矩阵，且方差 σ^2 满足某些条件（例如 $\sigma^2 \leq \frac{\log^4 n}{n}$），则最大奇异值 Δ 至少为 $2\sigma\sqrt{n}$。
   - **Davis-Kahan 定理**：
     假设两个对称矩阵 M 和 M' 有不同的特征值和特征向量。如果 $\Delta = M - M'$，则奇异值的界限可以控制特征向量的偏差。

> **意义**：这些定理帮助我们量化噪声（随机矩阵）对特征向量的影响。

Let $C = \mathbb{E}[C_M]$, and define the random matrix:

$$\Delta = C_M - C.$$

The matrix Δ is symmetric and random, with independent entries (except for symmetry). For each i, j, the expectation $\mathbb{E}[\Delta_{ij}] = 0$, and the variance $\mathrm{Var}(\Delta_{ij})$ is:

$$\mathrm{Var}(\Delta_{ij}) = \begin{cases} p(1-p), & \text{if } i = j, \\ p(1-p), & \text{if } i \neq j \text{ and } (i,j) \text{ in same community}, \\ q(1-q), & \text{if } i \neq j \text{ and } (i,j) \text{ in different communities}. \end{cases}$$

Note that $p(1-p)$, $q(1-q) \leq p + q$.

17. d-regular graph & Laplacian matrix (LG)

1. **d-正则图 G**：
   - 一个无向图，其中每个顶点都有 d 个邻居（度数为 d）。
   - 邻接矩阵 A_G 的定义：

$$A_G(i,j) = \begin{cases} 1, & \text{如果 } (i,j) \in E, \\ 0, & \text{否则.} \end{cases}$$

2. **Laplacian 矩阵 L_G**：定义为：

$$L_G(i,j) = \begin{cases} d, & \text{如果 } i = j, \\ -1, & \text{如果 } (i,j) \in E, \\ 0, & \text{否则.} \end{cases}$$

   其中：
   - 对角线元素是节点的度数 d；
   - 非对角线元素是边的存在性（负值）。

3. **性质**：
   - L_G 是对称矩阵。
   - 所有特征值都是非负的（重要结论）。

直观解释

- **对角元素** L_{ii}：表示顶点 i 的度数（与该顶点相连的边的条数）。
- **非对角元素** L_{ij}：如果 i 和 j 有边相连，值为 -1；如果无边相连，值为 0。
- **性质**：拉普拉斯矩阵是对称的（对无向图）且半正定。

重要性质

1. **特征值性质**：
   - 拉普拉斯矩阵的特征值非负。
   - 最小特征值总是 0，对应特征向量是全 1 向量。
   - 图的连通性和拉普拉斯矩阵的第二小特征值（称为**代数连通度**）密切相关。
2. **矩阵的行和为零**：
   - L 的每一行的元素和为 0。

矩阵形式：

$$L_{ij} = \begin{cases} \text{度数 (deg(i))}, & \text{如果 } i = j \\ -1, & \text{如果 } i \neq j \text{ 且 } (i,j) \in E \\ 0, & \text{如果 } i \neq j \text{ 且 } (i,j) \notin E \end{cases}$$

示例：一个简单的图

假设 G 是一个有 4 个顶点的图，边集为 $E = \{(1,2),(2,3),(3,4),(4,1)\}$。这是一个四边形。

1. **邻接矩阵 A**：

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

2. **度矩阵 D**：

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

3. **拉普拉斯矩阵 $L = D - A$**：

$$L = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

L : symmetric & PSD $\rightarrow x^T L x \geq 0$

$L v = \lambda v$

L_G 是 PSD $\rightarrow$ all eigenvalues of $L_G \geq 0$.

性质 1：L_G 是 PSD 矩阵（半正定矩阵）。

证明：

1. 考虑任意向量 x，定义 $x^T L_G x$：

$$x^T L_G x = \sum_{e \in E} L_e x,$$

   其中 L_e 是与边 $e = (u,v)$ 相关的拉普拉斯矩阵的贡献。

2. 展开后发现：

$$x^T L_G x = \sum_{e=(u,v) \in E} (x_u - x_v)^2,$$

   这是平方和，显然非负。

3. 因此，L_G 是半正定的，所有特征值都是非负的。

性质 2：L_G 的所有特征值非负：

$$0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

性质 3：$\lambda_1 = 0$，对应的特征向量是 $\mathbf{1}$。

- 当 $\mathbf{x} = \mathbf{1}$ 时，$x^T L_G x = 0$，因此 0 是 L_G 的特征值。
- 对应的特征向量是 $\mathbf{1}$（即每个元素都为 1 的向量）。

性质 4：$\lambda_2 = 0$ 的条件表示图是非连通的。

- 若图 G 是非连通的，可以将其划分为两个不相交的子集 S 和 $\bar{S}$，且 x 定义为：

$$x_n = \begin{cases} \frac{1}{|S|}, & \text{如果 } u \in S, \\ -\frac{1}{|\bar{S}|}, & \text{如果 } u \in \bar{S}. \end{cases}$$

  则对于任意边 $e = (u,v)$，$x_u = x_v$，从而 $x^T L_G x = 0$。

- 这表明存在另一个特征值 0（除了 λ_1），因此 $\lambda_2 = 0$。

性质 5：$\lambda_2 > 0$ 表示图是连通的。

- 如果 $\lambda_2 > 0$，则图是连通的，没有额外的特征向量对应于特征值 0。

Fact 2: If eigenvalues of L_G are $\lambda_1 \leq \ldots \leq \lambda_n$ then, $\lambda_1 = 0$. $\lambda_2 = 0$ iff G is disconnected.

- 最大特征值 λ_n 可以通过一个二次型的形式表示：

$$\lambda_n = \max_{x \neq 0} \frac{x^\top L_G x}{x^\top x} = \max_{x:\|x\|_2=1} x^\top L_G x$$

18. Positive Semi-definite (PSD)

1. PSD 矩阵的定义

一个 $n \times n$ 的对称矩阵 A 被称为 PSD 矩阵，如果对于任意非零向量 $x \in \mathbb{R}^n$，满足以下条件：

$$x^T A x \geq 0.$$

如果严格满足 $x^T A x > 0$（对所有 $x \neq 0$），则称为 **正定矩阵（Positive Definite Matrix, PD 矩阵）**。

---

2. PSD 矩阵的关键性质

1. **对称性**：PSD 矩阵必须是对称矩阵（即 $A = A^T$）。
2. **特征值非负**：PSD 矩阵的所有特征值 λ 都是非负的（即 $\lambda \geq 0$）。
3. **二次型非负**：对任意向量 x，二次型 $x^T A x \geq 0$。

19.

$$x^\top L_G x = x^\top \left(\sum_e L_e\right) x = \sum_e x^\top L_e x = \sum_{e=(u,v)} (x_u - x_v)^2$$

You can go over all the edges of the graph and look at the Laplacian for the edges but the edge Laplacian is this matrix:

$$L_e(i,j) = \begin{cases} 1 & \text{if} \quad i=j=u \text{ or } i=j=v \\ -1 & \text{if} \quad i=u, j=v \text{ or } i=v, j=u \\ 0 & \text{otherwise} \end{cases}$$

$$x^\top L_e x = \sum_{i,j} L_e(i,j) x_i x_j = x_u^2 + x_v^2 - 2x_u x_v = (x_u - x_v)^2$$

$$x^\top L_G x = x^\top \left(\sum_e L_e\right) x = \sum_e x^\top L_e x = \sum_{e=(u,v)} (x_u - x_v)^2 \le \sum_e 2\left(x_u^2 + x_v^2\right) = 2d\left(\sum_{u \in V} x_u^2\right)$$

$\lambda_n = 2d$ even cycles

Conclusion: The condition $\lambda_n = 2d$ implies that G has a bipartite component.

Corollary

If $\lambda_2 \neq 0$ and $\lambda_n = 2d$, $\Leftrightarrow$ then G is bipartite.
Here, $\lambda_2 \neq 0$ implies that the graph is connected (i.e., has only one connected component).

20. Max Cut & Conductance

- 在图 G 中，将顶点集合 V 划分为两个部分 S 和 $\bar{S} = V \setminus S$（即 S 的补集），我们关心两个集合之间的“边的流通情况”。
- **导通性（Conductance）**是一种衡量图划分“稀疏程度”的指标，它反映了划分后，跨越 S 和 $\bar{S}$ 的边占集合 S 的边总量的比例。

公式定义：

$$\Phi(S) = \frac{E(S,\bar{S})}{\mathrm{Vol}(S)} = \frac{E(S,\bar{S})}{d|S|}$$

其中：

- $\Phi(S)$：表示集合 S 的导通性。
- $E(S,\bar{S})$：跨越 S 和 $\bar{S}$ 的边数。
- $\mathrm{Vol}(S) = d|S|$：集合 S 的总边数（注意这里假设 G 是 d-正则图，节点度数固定为 d）。

Let $\lambda_2(G)$ be the second smallest eigenvalue of L_G.

Theorem:

$$\frac{\lambda_2(G)}{d} \le 2\Phi(G) \le O\left(\sqrt{\frac{\lambda_2(G)}{d}}\right)$$

$0 < \frac{\lambda_2(G)}{d} < 2$. Since $\lambda_2(G)$ can be calculated, an approximation of $\Phi(G)$ can be understood.

这说明：

- **下界**：导通性 $\Phi(G)$ 的最小值至少和 $\frac{\lambda_2(G)}{d}$ 成正比。
- **上界**：导通性 $\Phi(G)$ 的最小值最多和 $O\left(\sqrt{\frac{\lambda_2(G)}{d}}\right)$ 成正比。

3. Relaxed Goal（松弛目标）

由于精确求解 $\Phi(G)$ 很困难，松弛目标是找到集合 S，使得：

$$\Phi(S) \le 1.01\Phi(G)$$

即，找到一个“近似解”，其导通性最多比最优解高出 1%。

近似算法：

- **ARV 算法（Arora-Rao-Vazirani）**：
  - 输出集合 S，满足：
  $$\Phi(S) = O(\sqrt{\log n})\Phi(G)$$
  - 当 $\Phi(G) < \log n$ 时，该算法表现较好。
- **Cheeger 不等式算法：**
  - 输出集合 S，满足：
  $$\Phi(S) = O(\sqrt{\Phi(G)})$$
  - 当 $\Phi(G) > \log n$ 时，该算法效果较优。

给定一个**d-正则图** $G(V,E)$，目标是找到一个顶点集 S，使得：

- $|S| \le \frac{n}{2}$，即集合 S 的大小不超过总顶点数的一半；
- 最小化一个指标 $\Phi(S)$，定义为：

$$\Phi(S) = \frac{\text{穿过切割的边的数量}}{\text{集合 } S \text{ 的大小 } |S|}$$

sparsest cut: $\lambda_2 \le 2\Phi(G) = O(\Phi(G))$

21. convex set:

1. 关于凸集（Convex Set）的定义

一个集合 $A \subseteq \mathbb{R}^n$ 被称为**凸集**，当且仅当对于集合中的任意两个点 $x, y \in A$，其连接线上的任意一点（即 $\lambda x + (1-\lambda)y$ 对于 $\lambda \in [0,1]$）也属于该集合 A。公式定义为：

$$x, y \in A \implies \lambda x + (1-\lambda)y \in A \quad (\lambda \in [0,1])$$

第一步：证明 Δ 是凸集

集合 Δ 是所有 $\{z : z = \sum_{i=1}^m \lambda_i x_i\}$ 形成的凸包，满足权重 $\lambda_i \ge 0$ 且 $\sum_{i=1}^m \lambda_i = 1$。

1. 设 $z_1, z_2 \in \Delta$，则存在权重 $\{\lambda_i^1\}$ 和 $\{\lambda_i^2\}$，使得：
$$z_1 = \sum_{i=1}^m \lambda_i^1 x_i, \quad z_2 = \sum_{i=1}^m \lambda_i^2 x_i.$$
2. 对于任意 $\theta \in [0,1]$，定义 $z = \theta z_1 + (1-\theta) z_2$，则：
$$z = \theta \sum_{i=1}^m \lambda_i^1 x_i + (1-\theta)\sum_{i=1}^m \lambda_i^2 x_i.$$
3. 将权重合并为 $\lambda_i = \theta\lambda_i^1 + (1-\theta)\lambda_i^2$，有：
$$z = \sum_{i=1}^m \lambda_i x_i.$$
4. 检查权重是否满足条件：
   - 非负性：$\lambda_i = \theta\lambda_i^1 + (1-\theta)\lambda_i^2 \ge 0$；
   - 归一性：$\sum_{i=1}^m \lambda_i = \theta\sum_{i=1}^m \lambda_i^1 + (1-\theta)\sum_{i=1}^m \lambda_i^2 = 1$。

因此，Δ 是凸集。

假设：

- x^* 和 y^* 是线性规划的可行解。
- $\lambda x^* + (1-\lambda)y^*$ 是它们之间的线性组合，且 $\lambda \in [0,1]$。

证明：$\lambda x^* + (1-\lambda)y^*$ 也是可行解。

计算：

$$a^{(1)} \cdot (\lambda x^* + (1-\lambda)y^*) = \lambda(a^{(1)} \cdot x^*) + (1-\lambda)(a^{(1)} \cdot y^*)$$

由于 x^* 和 y^* 是可行解，满足 $a^{(1)} \cdot x^* \ge b_1$ 和 $a^{(1)} \cdot y^* \ge b_1$，因此：

$$\lambda a^{(1)} \cdot x^* + (1-\lambda)a^{(1)} \cdot y^* \ge \lambda b_1 + (1-\lambda)b_1 = b_1$$

类似的，所有约束 $a^{(i)} \cdot x \ge b_i$ 对 $i = 1, 2, \ldots, m$ 都成立。因此，$\lambda x^* + (1-\lambda)y^*$ 是可行解。

4. 顶点解（Vertex Solution）

定义： 在凸集中，一个点 $x \in C$ 是顶点（Vertex），如果对于 $\lambda \in [0,1]$，$x = \lambda y + (1-\lambda)z$（且 $y, z \in C$）时，总有 $y = x$ 或 $z = x$。

---

5. 线性规划的最优解出现在顶点

结论： 对于线性规划问题 $\min c \cdot x$，最优解要么出现在无界区域，要么出现在顶点。

证明要点：

- 如果最优解不是顶点，则可以表示为两个不同点 y^* 和 z^* 的线性组合：
$$x^* = \lambda y^* + (1-\lambda)z^*, \quad \lambda \in [0,1]$$
- $c \cdot x^* = \lambda c \cdot y^* + (1-\lambda)c \cdot z^*$。
- 如果 x^* 是最优解，y^* 或 z^* 至少有一个达到相同的最优值。因此，问题可以简化为检查凸集的顶点。

22. Convex function:

- **凸函数**：
  函数 f 是凸函数，当且仅当对于任意 $x, y \in \mathrm{dom}(f)$ 和 $\theta \in [0,1]$，有

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

  凸函数的直观解释是，其图像不会在连接任意两点的直线上方。

Definition of Convexity for a Function: A function $g(x)$ is convex if, for any $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in [0,1]$,

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2).$$

23. Maximum Flow & SVM

考虑一个有向图 $G = (V, E)$，其中每条边 $e \in E$ 都有一个容量 $c_e \geq 0$，表示这条边可以通过的最大流量。此外，指定两个顶点：**源点** s 和 **汇点** t。

目标是从源点 s 向汇点 t 发送最大流量，同时满足容量约束和流守恒约束。

约束条件

1. **非负性约束**：所有边上的流量 $x_e \geq 0$，即流量不能为负。
2. **容量约束**：$x_e \leq c_e$，即每条边上的流量不能超过其容量。
3. **流守恒约束**：对于除源点 s 和汇点 t 外的所有顶点 v，进入 v 的流量等于离开 v 的流量。即：

$$\sum_{e=(u,v)} x_e = \sum_{e=(v,w)} x_e.$$

目标

在满足上述约束的前提下，**最大化从源点到汇点的总流量**：

$$\max \sum_{e \in E} x_e,$$

SVM 是一种用于分类和回归分析的机器学习算法。这里简要提到 SVM 的两个模型。

特征和标签

- **特征 (Attributes)**：每个样本有 n 个特征，表示为：

$$x^{(1)}, x^{(2)}, \ldots, x^{(m)} \in \mathbb{R}^n.$$

  例如：房子的面积、卧室数量等。
- **标签 (Labels)**：对应的目标值或类别为：

$$a^{(1)}, a^{(2)}, \ldots, a^{(m)} \in \mathbb{R}.$$

模型 0

- 假设：

$$a^{(i)} = w \cdot x^{(i)},$$

  其中 w 是未知向量（分类模型中的权重向量）。
- 如果 w 存在，则可以通过**高斯消元法 (Gaussian Elimination)** 来求解。

模型 1

将特征表示为一个矩阵：

优化目标

我们需要评估向量 $a \in \mathbb{R}^m$ 与由 $b^{(1)}, b^{(2)}, \ldots, b^{(m)}$ 张成的子空间之间的接近程度。

目标是最小化：

$$\|a - \Sigma w b^{(i)}\|_2,$$

即向量 a 到张成子空间的 L_2 范数距离。

进一步扩展到误差建模

为了引入误差的约束，目标误差被定义为 ξ，并约束每个样本的误差满足：

$$\|a^{(i)} - w \cdot x^{(i)}\| \leq \xi \cdot a^{(i)} \quad \text{for } i \leq m.$$

这意味着我们希望每个数据点 $a^{(i)}$ 的预测值 $w \cdot x^{(i)}$ 尽可能接近真实值，同时限制误差的比例不超过 $\xi \cdot a^{(i)}$。

线性化约束

通过分解，可以将误差约束改写为线性约束形式：

$$-\xi^{(i)} \cdot a^{(i)} \leq a^{(i)} - w \cdot x^{(i)} \leq \xi^{(i)} \cdot a^{(i)},$$

并以最小化 ξ 作为目标。

24. Approximation Algorithm & Vertex Cover

对于难以精确解决的优化问题（如 NP-hard 问题），使用近似算法是一个常见策略。目标是找到接近最优解的可行解。

- **定义**：给定一个最小化问题 C，算法 A 被称为 α-近似算法，如果对于任何实例 I，其输出满足：$A(I) \leq \alpha \cdot OPT(I)$ 其中：
  - $OPT(I)$ 是最优解的值，
  - $\alpha \geq 1$ 是近似比，表示算法结果与最优解的偏差。

给定一个无向图，目标是找到一个顶点的子集，使得图中的每条边至少有一个端点在该子集中。

---

1. 问题定义

- **顶点覆盖问题**：目标是找到一个最小顶点集，使得图中每条边的至少一个端点在该集合中。
- **目标**：设计一个 2-近似算法，即找到的顶点覆盖的大小 S 满足：

$$|S| \leq 2 \cdot OPT$$

  其中 OPT 是问题的最优解。

2. 原始问题的线性规划形式 (OPT∗)

- 变量 x_u：表示顶点 u 是否属于顶点覆盖，取值为 0 或 1。
- 约束条件：
  - $x_u + x_v \geq 1$：每条边 (u, v) 至少有一个端点被覆盖。
- 目标函数：

$$\min \sum_{u \in V} x_u$$

  即最小化顶点覆盖的大小。

---

3. 松弛后的线性规划 (LP∗∗)

通过将变量 $x_u \in \{0,1\}$ 松弛为 $0 \leq x_u \leq 1$，问题变为：

$$\min \sum_{u \in V} x_u$$

同时满足：

- $x_u + x_v \geq 1$：每条边至少覆盖一个端点。

Claim 1: S 是图 G 的顶点覆盖。

证明：对于任意边 (u, v)，线性规划的解必须满足 $x_u + x_v \geq 1$。因此，至少有一个 $x_u \geq 0.5$ 或 $x_v \geq 0.5$，故 S 中包含边的至少一个端点。

1. 构造的顶点覆盖 S

根据算法，顶点覆盖 S 是通过如下规则构造的：

$$S = \{u \in V : x_u \geq 0.5\},$$

其中 x_u 是松弛线性规划的最优解。

2. 分析 $|S|$ 和 $\sum_{u \in V} x_u$

根据定义，$|S|$ 可以写成：

$$|S| = \sum_{u \in V} 1(u \in S),$$

其中 $1(u \in S)$ 是一个指示函数，当 $u \in S$ 时等于 1，否则等于 0。

3. 取整的关键性质

当 $u \in S$ 时，根据规则 $x_u \geq 0.5$。因此，每个 $u \in S$，我们有：

$$\frac{1(u \in S)}{x_u} \leq \frac{1}{0.5} = 2.$$

4. 推导不等式

将上述性质应用到 $|S|$ 的表达式中，我们可以写成：

$$|S| = \sum_{u \in V} 1(u \in S) = \sum_{u \in V} \frac{1(u \in S)}{x_u} \cdot x_u.$$

将每一项 $\frac{1(u \in S)}{x_u}$ 的值替换为 ≤ 2，得到：

$$|S| \leq \sum_{u \in V} 2 \cdot x_u.$$

因此：

$$|S| \leq 2 \cdot \sum_{u \in V} x_u.$$

---

5. 结论

由于 $\sum_{u \in V} x_u$ 是松弛线性规划问题 (∗∗) 的最优目标值（即 $\mathrm{OPT\ of\ the\ LP}$），我们最终得到：

$$|S| \leq 2 \cdot (\mathrm{OPT\ of\ the\ LP}),$$

从而证明了 Claim 2。

k-klique:
Best vertex cover: k-1
Best fractional solution: $\frac{k}{2}$

25. Set Cover

- **输入**：
  - 一个全集 $U = \{a_1, a_2, \ldots, a_n\}$，包含 n 个元素。
  - 一组子集 $S_1, S_2, \ldots, S_m \subseteq U$。
  - 子集的并集必须覆盖全集：$\bigcup_{i \in [m]} S_i = U$。
- **目标**：找到一个最小的子集索引集合 $I \subseteq [m]$，使得：

$$\bigcup_{i \in I} S_i = U$$

即，选择尽可能少的子集覆盖全集。

* 1. 决策变量：
   - $x_j \in \{0,1\}, \forall 1 \le j \le m$
     - $x_j = 1$ 表示选择了集合 S_j；
     - $x_j = 0$ 表示未选择集合 S_j。
2. 约束条件：
   - $\forall i \in [n], \sum_{j: i \in S_j} x_j \ge 1$
     - 每个元素 i 必须被至少一个被选择的集合覆盖。
3. 目标函数：
   - 最小化 $\sum_{j=1}^{m} x_j$
     - 最小化选择集合的数量。

23.2.3 线性规划松弛（LP Relaxation）

为了求解集合覆盖问题，可以将整数线性规划（ILP）转化为线性规划（LP），这是通过松弛约束来实现的。我们将该线性规划表示为 $(**)$。

表述：

1. **决策变量松弛**：
   - 在 ILP 中，$x_j \in \{0,1\}$，表示集合 S_j 是否被选中；
   - 在 LP 中，松弛为 $0 \le x_j \le 1$，允许 x_j 是一个连续值。
2. **约束条件**：
   - $\forall i \in [n], \sum_{j: i \in S_j} x_j \ge 1$
     - 与 ILP 相同，每个元素 i 必须被至少一个集合覆盖。
3. **目标函数**：
   - 最小化 $\sum_{j=1}^{m} x_j$
     - 尽可能减少选中的集合数量。

$$\forall j \in [n], \quad \sum_{i: j \in S_i} x_i \ge 1.$$

个约束条件的含义是什么?

表示形式：

- 这个不等式对全集中的每个元素 $j \in U$ 都要成立。
- $i : j \in S_i$ 表示所有包含元素 j 的子集 S_i。
- x_i 是指是否选择了子集 S_i：
  - 如果 $x_i = 1$，表示子集 S_i 被选中；
  - 如果 $x_i = 0$，表示子集 S_i 没有被选中。

解释：这个约束的意思是：**全集 U 中的每个元素 j 必须至少被一个被选中的子集覆盖。**

定理 23.1: 存在一个 $O(\log n)$ 逼近算法来解决集合覆盖问题，这种逼近可以改进为 $\ln n + \Theta(1)$

Claim: Objective value of (**) $\le$ objective value of (*). Since (**) is a relaxation of (*), any solution of (*) satisfies the requirements of (**), which means that the objective value of (**) is more optimized than (*), thus smaller in our case.

26. Randomized Approximation

Fact: $1 - x \le e^{-x}$ for $x \in [0,1]$. This will be used in later steps.

1. **随机选择集合**：
   - 对于每个集合 S_j，以概率 x_j 将其加入集合 C（这里的 x_j 是从线性规划（LP）解中得到的）。
2. **未被覆盖的概率**：
   - 一个元素 i 未被覆盖的概率为：

$$\prod_{j: i \in S_j} (1 - x_j)$$

   - 基于 $1 - x \le e^{-x}$ 的不等式，这个概率可以近似为：

$$\prod_{j: i \in S_j} e^{-x_j} = e^{-\sum_{j: i \in S_j} x_j} \le \frac{1}{e}$$

   - 这表示每个元素 i 被漏掉的概率最多是 $\frac{1}{e}$。

Let the number of sets added to C in round i be y_i. The expectation of y_i is:

$$\begin{aligned}\mathbb{E}[y_i] &= \sum_{j=1}^{m} \mathbb{E}[\mathbb{1}[S_j \text{ was added to } C]] \\ &= \sum_{j=1}^{m} x_j \\ &= L\end{aligned}$$

Since $|C| \le y_1 + \cdots y_T$, $\mathbb{E}[|C|] \le \mathbb{E}[y_1] + \cdots + \mathbb{E}[y_T] = TL$.

第一步：单轮覆盖概率分析

公式：

$$\prod_{j: i \in S_j} (1 - x_j) \le e^{-\sum_{j: i \in S_j} x_j} \le \frac{1}{e}$$

- 解释：
  - 对于元素 i，在一轮中它未被覆盖的概率是 $\prod_{j: i \in S_j}(1 - x_j)$。
  - 利用 $1 - x \le e^{-x}$，将累乘松弛为累加的指数形式：

$$\prod_{j: i \in S_j} (1 - x_j) \le e^{-\sum_{j: i \in S_j} x_j}$$

  - 根据线性规划的约束条件 $\sum_{j: i \in S_j} x_j \ge 1$，进一步得到：

$$e^{-\sum_{j: i \in S_j} x_j} \le e^{-1} = \frac{1}{e}$$

- 结论：在单轮中，元素 i 未被覆盖的概率最多是 $\frac{1}{e}$。

第二步：多轮覆盖概率分析

公式：

$$\text{在 } T \text{ 轮后，元素 } i \text{ 未被覆盖的概率为：} \ e^{-T}$$

- 假设每轮都是独立的，T 轮后，元素 i 未被覆盖的概率是每一轮未被覆盖概率的乘积：

$$\prod_{\text{轮次}} e^{-1} = e^{-T}$$

- 设置 $T = \ln n + 10$，未覆盖概率变为：

$$e^{-(\ln n + 10)} = e^{-\ln n} \cdot e^{-10} = \frac{1}{n} \cdot e^{-10}$$

- 结论：单个元素 i 未被覆盖的概率极小，只有 $\frac{e^{-10}}{n}$。

第三步：全集未覆盖概率分析

公式：

$$\text{全集 } U \text{ 中存在某个 } i \text{ 未被覆盖的概率：} \ n \cdot \frac{e^{-10}}{n} = e^{-10}$$

- 使用联合概率界限（Union Bound）：
  - 对于全集 U，有 n 个元素；
  - 某个元素未被覆盖的概率为 $\frac{e^{-10}}{n}$；
  - 那么全集中至少有一个未被覆盖的概率最多是 $n \cdot \frac{e^{-10}}{n} = e^{-10}$。
- 结论：$T = \ln n + 10$ 轮后，全集 U 被完全覆盖的概率为：

$$1 - e^{-10} \ (\text{接近 } 99.999\%)$$

第一步：用 Markov 不等式限制集合大小

Markov 不等式告诉我们，对于非负随机变量 X，

$$\Pr(X \ge a) \le \frac{\mathbb{E}[X]}{a}.$$

将其应用到集合大小 $|C|$ 上：

1. **随机变量**：
   - $X = |C|$，表示集合 C 的实际大小。
2. **期望值**：
   - $\mathbb{E}[|C|] = T \cdot L$。
3. **阈值**：
   - 选择 $a = 10 \cdot T \cdot L$，即我们希望控制集合大小不会超过 $10 \cdot T \cdot L$。
4. **结果**：
   - 根据 Markov 不等式：

$$\Pr(|C| \ge 10 \cdot T \cdot L) \le \frac{\mathbb{E}[|C|]}{10 \cdot T \cdot L} = \frac{T \cdot L}{10 \cdot T \cdot L} = \frac{1}{10}.$$

- **解释**：集合 C 的大小超过 $10 \cdot T \cdot L$ 的概率最多是 $\frac{1}{10}$。换句话说，有 90% 的概率，集合 $|C| \le 10 \cdot T \cdot L$。

Suppose we set $T = \ln n + 10$, then the probability that i is not covered in T rounds $\le e^{-(\ln n + 10)} \le \frac{e^{-10}}{n}$.
As a result, the probability that $\exists\, i \in [n]$ not covered in T rounds $\le n \cdot \frac{e^{-10}}{n} = e^{-10}$ (by union bound)
$\Rightarrow$ w.p. $1 - e^{-10}$, every element i is covered.

By Markov's Inequality, w.p. $\frac{9}{10}$:

$$\begin{aligned}|C| &\le 10 \cdot T \cdot L \\ &= O(L \log n)\end{aligned}$$

$\therefore$ w.p. $1 - e^{-10} - \frac{1}{10}$, both events happen. The probability evaluates to $\ge 89.9\%$.

8. Markov's inequaltiy

Suppose Z is a non-negative random variable, then $\forall t \ge 1$,
$\Pr[Z \ge t \cdot \mathbb{E}[Z]] \le \frac{1}{t}$

HW 3:

1. [3 + 3 points] (i) Suppose A is a symmetric matrix. Then prove that all the eigenvalues of A are non-negative if and only if $x^T Ax \geq 0$ for every $x \in \mathbb{R}^n$. (Remark: such a matrix A is called a psd matrix).

(ii) Suppose $A \in \mathbb{R}^{n\times n}$ is a psd matrix. Prove that for any matrix $B \in \mathbb{R}^{m\times n}$, BAB^T is also psd.

2. [5 points] Suppose A is a square matrix of size $n \times n$ (not necessarily symmetric). Let the SVD of A be given by $\sum_{i=1}^n \sigma_i u_i v_i^T$ where $\sigma_i \neq 0$ (for all i). Prove the the inverse of A is given by $\sum_{i=1}^n \sigma_i^{-1} v_i u_i^T$. (Recall that the inverse of a square matrix A is the unique matrix B such that $BA = AB = I$).

3. [3 points] (a) Suppose A is symmetric matrix. We know from standard linear algebra that $\mathsf{Tr}(A) = \sum_{i=1}^n \lambda_i$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Prove that $\mathsf{Tr}(A^k) = \sum_{i=1}^n \lambda_1^k + \ldots + \lambda_n^k$.

(b) [5 points] Consider a random matrix $A \in \mathbb{R}^n$ defined as follows: entry $A(i,j)$ (where $i \leq j$) is set to ± 1 with probability $1/2$. If $i > j$, then $A(i,j)$ is set to be equal to $A(j,i)$. Compute $\mathbf{E}[\mathsf{Tr}(A^4)]$ and use that to prove that $\max_i |\lambda_i| = O(n^{3/4})$ with probability 0.99.

Hint: For part (b), write the expression for $\mathsf{Tr}(A^4)$ in terms of the entries of A and then use simple linearity of expectation.

1

1 Question 1

1. Let's first assume all the eigenvalues of A are non-negative. Because we know A is symmetric, we know it can be diagonalized to $A = U\Lambda U^\top$ where U is an orthogonal matrix and Λ is a diagonal matrix which contains eigenvalues $\lambda_i \geq 0$. We also know $U^\top U = I$.

   Then, for any vector $x \in \mathbb{R}^n$, we have:

   $$x^\top Ax = x^\top U\Lambda U^\top x = (U^\top x)^\top \Lambda (U^\top x)$$

   We set $y = Q^\top x$, and the equation would become $(U^\top x)^\top \Lambda (U^\top x) = y^\top \Lambda y$

   Because we know Λ is a diagonal matrix which contains eigenvalues $\lambda_i \geq 0$, we know all of its entries are non-negative. Thus, we can deduce:

   $$x^\top Ax = y^\top \Lambda y = \sum_{i=1}^n \lambda_i y_i^2 \geq 0$$
   $$x^\top Ax \geq 0$$

   Proof by contradiction: if A has a negative eigenvalue $\lambda_x < 0$, and its eigenvector is $v_x \neq 0$, we have:

   $$v_x^\top Av_x = v_x^\top(\lambda_x v_x) = \lambda_x v_x^\top v_x = \lambda_x \|v_x\|^2 < 0$$

   This contradicts with the conclusion $v^\top Av \geq 0$, so all the eigenvalues of A must be non-negative.

2. Suppose we have a random vector $x \in \mathbf{R}^m$, we want to put it in the matrix $BAB^\top$:

   $$x^\top(BAB^\top)x = (B^\top x)^\top A(B^\top x)$$

   Suppose we have another vector $y = B^\top x \in \mathbf{R}^n$, because we know A is psd, $y^\top Ay \geq 0$. Thus, we have:

   $$x^\top(BAB^\top)x = y^\top Ay \geq 0$$
   $$x^\top(BAB^\top)x \geq 0$$

2 This holds for all $x \in \mathbb{R}^m$. Therefore, $BAB^\top$ is psd.

3

2 Question 2

$$A = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

Alternatively, we can rewrite in using the following form:

$$A = U\Sigma V^\top$$

where U is an orthogonal matrix made up of $[u_1, u_2, ..., u_n]$, V is an orthogonal matrix made up of $[v_1, v_2, ..., v_n]$, and Σ is a diagonal matrix made up of $[\sigma_1, \sigma_2, ..., \sigma_n]$. Because all $\sigma_i \neq 0$, Σ is invertible with its inverse:

$$\Sigma^{-1} = \mathrm{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_n^{-1})$$

The inverse of A is then $A^{-1} = V\Sigma^{-1}U^\top$. We also know $\Sigma^{-1} = \sum_{i=1}^n \sigma_i^{-1} e_i e_i^\top$ where e_i are the standard basis vectors ($Ve_i = v_i$, and $(Ue_i)^\top = u_i^\top$). If we expand it, we would get:

$$\begin{aligned}
A^{-1} &= V\Sigma^{-1}U^\top \\
&= V(\sum_{i=1}^n \sigma_i^{-1} e_i e_i^\top)U^\top \\
&= \sum_{i=1}^n \sigma_i^{-1} V e_i e_i^\top U^\top \\
&= \sum_{i=1}^n \sigma_i^{-1} v_i u_i^\top
\end{aligned}$$

Therefore, the inverse of A is given by $\sum_{i=1}^n \sigma_i^{-1} v_i u_i^\top$.

3 Question 3

1. Let's first assume all the eigenvalues of A are non-negative. Given that A is symmetric, we know it can be diagonalized to $A = U\Lambda U^\top$ where U is an orthogonal matrix and Λ is a diagonal matrix which contains eigenvalues $\lambda_i \geq 0$. We also know $U^\top U = I$.

   $$A^k = (U\Lambda U^\top)^k = U\Lambda U^\top \cdots U\Lambda U^\top = U\Lambda^k U^\top$$

   Λ^k is a diagonal matrix with entries $\lambda_1^k \ldots \lambda_n^k$

   The trace of a matrix is invariant under orthogonal matrix.

   $$\mathrm{Tr}(A^k) = \mathrm{Tr}(U\Lambda^k U^\top) = \mathrm{Tr}(\Lambda^k)$$

   Therefore $\mathrm{Tr}(A^k) = \sum_i^n \lambda_i^k$

4

2. We first deduce the expression of $Tr(A^4)$ in terms of the entries of A:

   $$\mathrm{Tr}(A^4) = \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \sum_{l=1}^n A_{ij}A_{jk}A_{kl}A_{li}$$

   The index is a close loop, since the trace is the sum of the diagonal position.

   Based on the linearity of expectation:

   $$\mathbb{E}[\mathrm{Tr}(A^4)] = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \mathbb{E}[A_{ij}A_{jk}A_{kl}A_{li}]$$

   Note that A is symmetric and the number on each entry is randomly chosen and has an expectation of 0. Therefore the expectation of each entry is non-zero when every entry appears an even number of times.

   There are 3 scenarios:

   1. all the 4 indices are the same, i.e $i = j = k = l$. The expectation of this term is:

   $$\sum_{i=1}^n \mathbb{E}[A_{ii}^4] = n$$

   2. 3 indices are the same, e.g. $j = k = l$. This pattern appears 4 times. The expectation of this term is:

   $$4\sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathbb{E}[A_{jj}^2 A_{ij}^2] = 4n^2 - 4n$$

   3. 2 indices are the same, i.e. $i = k, j = l$ and $i = j, k = l$ and $i = l, j = k$. This pattern appears 3 times. The expectation of this term is:

   $$\sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathbb{E}[A_{ij}^4] + 2\sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathbb{E}[A_{jj}^2 A_{ij}^2] = 3n^2 - 3n$$

   Therefore the explicit expression of expectation is:

   $$\mathbb{E}[\mathrm{Tr}(A^4)] = 7n^2 - 6n$$

   By Markov's inequality, let $Z = \mathrm{Tr}(A^4), t = \frac{n^4}{7n^2 - 6n}$:

   $$\Pr[Z \geq t\mathbb{E}[Z]] \leq \frac{1}{t}$$
   $$\Pr[Z \geq n^4] \leq \frac{7n - 6}{n^3}$$

6

Noted that:

$$Z = \mathrm{Tr}(A^4) = \sum_i^n \lambda_i^4 \leq n \cdot \max_i |\lambda_i|^4$$

$\max_i |\lambda_i|^4$ is also a random variable, so we can determine its inequality by the above Markov's inequality:

$$\Pr[n \cdot \max_i |\lambda_i|^4 \geq n^4] \leq \frac{7n - 6}{n^3}$$
$$\Pr[\max_i |\lambda_i| \geq n^{\frac{3}{4}}] \leq \frac{7n - 6}{n^3}$$

When n is large enough, without the loss of generality, let $n = 100$, $\Pr[\max_i |\lambda_i| \geq n^{\frac{3}{4}}] \leq 0.0007$

Therefore $\max_i |\lambda_i| = O(n^{3/4})$ with probability at least 0.99.

HW4:

1. [5 points] Let G be a d-regular graph and let L_G be its Laplacian. Let the eigenvalues of L_G be $\lambda_1 \leq \ldots \leq \lambda_n$. Prove that $\lambda_k = 0$ iff the Graph G has at least k connected components.

2. (i) [3 points] A matrix $P \in \mathbb{R}^{n\times n}$ is said to be a permutation matrix if there is a permutation $\sigma : [n] \to [n]$ such that $P(i,j) = 1$ if and only if $\sigma(i) = j$. Prove that if $Av = \lambda v$ for some vector v, then, Pv is an eigenvector of PAP^{-1} with eigenvalue λ.

(b) [3 points] For a permutation $\sigma : [n] \to [n]$, let P be the permutation matrix. Then, show that (i,j)th entry of PAP^{-1} is $A_{\sigma(i),\sigma(j)}$.

3. [4 points] (a) Suppose $A \in \mathbb{R}^{m\times n}$. Let the SVD of A be $\sum_i \sigma_i u_i v_i^T$. Prove that

$$\max_{x\in\mathbb{R}^m,\ \|x\|_2=1,\ y\in\mathbb{R}^n\ \|y\|_2=1} x^T Ay = \max_i \sigma_i.$$

[5 points] (b) For a rectangular matrix A, we often use $\|A\|_2$ to denote the maximum signular value of A. Prove the following inequalities:

$$\|A+B\|_2 \leq \|A\|_2 + \|B\|_2;$$

$$\|A \cdot B\|_2 \leq \|A\|_2 \cdot \|B\|_2.$$

4. [5 points] Suppose you have a vector $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$. Consider another vector u which encodes a set $S \subseteq [n]$ as follows: If $i \in S$, $u_i = 1/\sqrt{n}$. If $i \notin S$, $u_i = -1/\sqrt{n}$.

You're given v but not u. Furthermore, you're told that $\|v-u\|_2 \leq \epsilon$. Let us define $S' = \{i : v_i \geq 0\}$. Prove that $|S\Delta S'| = O(\epsilon^2 n)$.

1

1 Question 1

1. Let c be the number of connected components in G. Since we know the multiplicity of the eigenvalue zero of the Laplacian matrix L_G equals the number of connected componets in G, we can deduce that the multiplicity of the eigenvalue zero is c. This implies the smallest c eigenvalues are all zero:

$$\lambda_1 = \lambda_2 = \cdots = \lambda_c = 0$$

It obvious that $c \geq k$, so $\lambda_k = 0$. **Therefore, if G has at least k connected components, $\lambda_k = 0$.**

Secondly, let us assume $\lambda_k = 0$, this implies the multiplicity of the zero eigenvalue is at least k. Since we also know the multiplicity of the zero eigenvalue equals the number of connected components, G should have at least k connected components. **Therefore, if $\lambda_k = 0$, G should have at least k connected components.**

Thus, **$\lambda_k = 0$ iff the Graph G has at least k connected components.**

2 Question 2

1. Consider the expression $(PAP^{-1})(Pv)$, we first regroup this expression:

$$(PAP^{-1})(Pv) = (PA)(P^{-1}Pv)$$

Notice that $P^{-1}P = I$, where I is the identity matrix:

$$P^{-1}Pv = v$$

Thus, the expression simplifies to:

$$(PAP^{-1})(Pv) = (PA)v = P(Av)$$

Notice that v is the eigenvector of A:

$$Av = \lambda v$$

2

3

Substitute it to the former expression. Notice that λ is a scalar:

$$P(Av) = P(\lambda v) = \lambda(Pv)$$

Therefore, we have show that:

$$(PAP^{-1})(Pv) = \lambda(Pv)$$

This means Pv is an eigenvector of PAP^{-1} with eigenvalue λ

2. Consider the property of the inverse P^{-1}:

$$I(i,j) = \sum_{k=1}^{n} P(i,k)P^{-1}(k,j)$$

$P(i,k) = 1$ if and only if $k = \sigma(i)$. $I(i,j) = 1$ if and only if $i = j$. Thus, the variable k collapses to one term:

$$1 = P(i,\sigma(i))P^{-1}(\sigma(i), i)$$

Which means for every i:

$$P^{-1}(\sigma(i), i) = 1$$

Now consider the (i,j)th entry of PAP^{-1}:

$$PAP^{-1}(i,j) = \sum_{k=1}^{n}\sum_{l=1}^{n} P(i,k)A(k,l)P^{-1}(l,j)$$

$P(i,k) = 1$ if and only if $k = \sigma(i)$. Thus, the sum over k collapses to the term where $k = \sigma(i)$:

$$PAP^{-1}(i,j) = \sum_{l=1}^{n} A(\sigma(i), l)P^{-1}(l,j)$$

$P^{-1}(l,j) = 1$ if and only if $l = \sigma(j)$. Thus, the sum over l collapses to the term where $l = \sigma(j)$:

$$PAP^{-1}(i,j) = A(\sigma(i), \sigma(j))$$

Therefore the (i,j)th entry of PAP^{-1} is $A_{\sigma(i),\sigma(j)}$

3 Question 3

1. We know $\{u_i\}$ and $\{v_i\}$ form orthonormal bases, so unit vectors x and y can be expressed as follows:

$$x = \sum_i \alpha_i u_i, \quad y = \sum_i \beta_i v_i$$

We also know that $\sum_i \alpha_i^2 = 1$ and $\sum_i \beta_i^2 = 1$.

4

Then, we want to compute $x^T Ay$.

$$x^T Ay = \left(\sum_i \alpha_i u_i^T\right)\left(\sum_j \sigma_j u_j v_j^T\right)\left(\sum_k \beta_k v_k\right) = \sum_i \sigma_i \alpha_i \beta_i.$$

To maximize $x^T Ay$, we need to maximize $\sum_i \sigma_i \alpha_i \beta_i$ under the constraints we mentioned above.

By applying Cauchy-Schwarz inequality ($|\sum_i a_i b_i| \leq (\sum_i a_i^2)^{1/2}(\sum_i b_i^2)^{1/2}$) and set $a_i = \sigma_i^{1/2}\alpha_i$ and $b_i = \sigma_i^{1/2}\beta_i$, we get:

$$\sum_i \sigma_i \alpha_i \beta_i \leq \left(\sum_i \sigma_i \alpha_i^2\right)^{1/2} \left(\sum_i \sigma_i \beta_i^2\right)^{1/2}$$

Since σ_i is in decreasing order, $\sigma_1 = \max_i \sigma_i$ because it is the largest. The maximum value is achieved when $\alpha_1 = 1$, $\beta_1 = 1$, and all other α_i, β_i are zero.

Therefore:

$$\max_{\substack{x\in\mathbb{R}^m,\ \|x\|_2=1 \\ y\in\mathbb{R}^n,\ \|y\|_2=1}} x^T Ay = \sigma_1 = \max_i \sigma_i$$

$$\max_{\substack{x\in\mathbb{R}^m,\ \|x\|_2=1 \\ y\in\mathbb{R}^n,\ \|y\|_2=1}} x^T Ay = \max_i \sigma_i$$

2. (a) Operrator norm is as follows:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

Based on triangle inequality, we can deduce that for any x with $\|x\|_2 = 1$:

$$\|(A+B)x\|_2 \leq \|Ax\|_2 + \|Bx\|_2$$

If we divide both sides by $\|x\|_2$, the equation remains the same as $\|x\|_2 = 1$. Take the maximum over all x, we get:

$$\|A+B\|_2 = \max_{\|x\|_2=1} \|(A+B)x\|_2 \leq \max_{\|x\|_2=1} (\|Ax\|_2 + \|Bx\|_2) \leq \|A\|_2 + \|B\|_2$$

Therefore:

$$\|A+B\|_2 \leq \|A\|_2 + \|B\|_2$$

(b) We start by recalling the definition of the operator norm for a product of matrices:

$$\|AB\|_2 = \max_{\|x\|_2=1} \|ABx\|_2$$

Using the property of norms, we deduce that:

$$\|ABx\|_2 \leq \|A\|_2 \|Bx\|_2$$

5

Next, we bound $\|Bx\|_2$ by applying the operator norm of B. We also know $\|x\|_2 = 1$:

$$\|Bx\|_2 \leq \|B\|_2\|x\|_2 = \|B\|_2$$

Combine these inequalities, we get this:

$$\|ABx\|_2 \leq \|A\|_2\|B\|_2$$

Finally, taking the maximum over all x with $\|x\|_2 = 1$, we deduce:

$$\|AB\|_2 = \max_{\|x\|_2=1} \|ABx\|_2 \leq \|A\|_2\|B\|_2$$

Thus,

$$\|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$$

4 Question 4

1. Given that $\|v-u\|_2 \leq \epsilon$. This means:

$$\sum_{i=1}^{n}(v_i - u_i)^2 \leq \epsilon^2$$

$|S\Delta S'|$ denote the size of the symmetric difference between sets S and S', which is:

$$|S\Delta S'| = |(S \setminus S') \cup (S' \setminus S)|$$

Notice that $S = \{i : u_i \geq 0\}, S' = \{i : v_i \geq 0\}$. By definition, the size of the symmetric difference increases when:

$$u_i \geq 0, v_i < 0 \text{ or } u_i < 0, v_i \geq 0$$

1. when $u_i \geq 0, v_i < 0$ then $u_i = \frac{1}{\sqrt{n}}, v_i < 0$. The squared difference is:

$$(v_i - u_i)^2 \geq (\frac{1}{\sqrt{n}})^2 = \frac{1}{n}$$

2. when $u_i < 0, v_i \geq 0$ then $u_i = -\frac{1}{\sqrt{n}}, v_i \geq 0$. The squared difference is:

$$(v_i - u_i)^2 \geq (\frac{1}{\sqrt{n}})^2 = \frac{1}{n}$$

Let d be the size of the symmetric difference, i.e. $d = |S\Delta S'|$, then:

$$d \cdot \frac{1}{n} \leq \sum_{i \in S\Delta S'}(v_i - u_i)^2 \leq \sum_{i=1}^{n}(v_i - u_i)^2 \leq \epsilon^2$$

Thus:

$$d \leq \epsilon^2 n$$

This proves that $|S\Delta S'| = O(\epsilon^2 n)$

HW5:

1. You already know what it means for a set D to be convex. A function $f : D \to \mathbb{R}$ is said to be convex if for every $x, y \in D$ and $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

(a) [3 points] For $t \in \mathbb{R}$, prove that the set $D_t = \{x : x \in D \text{ and } f(x) \leq t\}$ is a convex set.

(b) [3 points] Let A be a positive semidefinite matrix. Then, for any $\theta \geq 0$, prove that $Z_{A,\theta} = \{x \in \mathbb{R}^n : x^T Ax \leq \theta\}$ is a convex set.

1 Question 1

1. We want to show D_t is convex, so based on the definition of convex set, we need to prove: for any $m, n \in D_t$ and $\lambda \in [0,1]$, point $z = \lambda m + (1-\lambda)n$ also belongs to D_t.

   We know $m, n \in D_t$, and for any $x \in D_t$, $f(x) \leq t$, so we can deduce: $f(m) \leq t$, and $f(n) \leq t$.

   We know f is a convex function, so based on the convexity of the function, we have:

$$f(z) = f(\lambda m + (1-\lambda)n) \leq \lambda f(m) + (1-\lambda)f(n) \leq \lambda t + (1-\lambda)t = t$$

   So,

$$f(z) \leq t$$

   This implies:

$$z \in D_t$$

   Thus, D_t is indeed a convex set.

2. We first want to prove function $f(x) = x^\top Ax$ is convex. That's to say, we want to prove:

   For any $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

$$f(\lambda x+(1-\lambda)y) = (\lambda x+(1-\lambda)y)^\top A(\lambda x+(1-\lambda)y) = \lambda^2 x^\top Ax+2\lambda(1-\lambda)x^\top Ay+(1-\lambda)^2 y^\top Ay$$

   Then, we compute $\lambda f(x) + (1-\lambda)f(y)$:

$$\lambda f(x) + (1-\lambda)f(y) = \lambda x^\top Ax + (1-\lambda)y^\top Ay$$

   Now, we take the difference of the two (which is $f(\lambda x+(1-\lambda)y)-[\lambda f(x)+(1-\lambda)f(y)]$):

$$\begin{aligned}\text{difference} &= [\lambda^2 x^\top Ax + 2\lambda(1-\lambda)x^\top Ay + (1-\lambda)^2 y^\top Ay] - [\lambda x^\top Ax + (1-\lambda)y^\top Ay] \\ &= \lambda^2 x^\top Ax - \lambda x^\top Ax + 2\lambda(1-\lambda)x^\top Ay - (1-\lambda)y^\top Ay \\ &= -\lambda(1-\lambda)x^\top Ax + 2\lambda(1-\lambda)x^\top Ay - \lambda(1-\lambda)y^\top Ay \\ &= \lambda(1-\lambda)\left[-x^\top Ax + 2x^\top Ay - y^\top Ay\right]\end{aligned}$$

We notice:

$$\text{difference} = \lambda(1-\lambda)\left[-x^\top Ax + 2x^\top Ay - y^\top Ay\right] = \lambda(1-\lambda)\left[-(x-y)^\top A(x-y)\right]$$

Because we know A is psd and symmetric, $(x-y)^\top A(x-y) \geq 0$. Thus:

$$\text{difference} \leq 0$$

Thus:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

Therefore, $f(x) = x^\top Ax$ is convex.

Secondly, we want to prove $Z_{A,\theta}$ is a convex set:

For $m, n \in Z_{A,\theta}$, we have $m^T Am \leq \theta$ and $n^T An \leq \theta$.

For $\lambda \in [0,1]$, we have:

$$(\lambda m + (1-\lambda)n)^T A(\lambda m + (1-\lambda)n) \leq \lambda m^T Am + (1-\lambda)n^T An \leq \lambda\theta + (1-\lambda)\theta = \theta$$

This implies:

$$\lambda m + (1-\lambda)n \in Z_{A,\theta}$$

Thus, $Z_{A,\theta}$ is a convex set.