Saadullah Khan

**Problem 1** (2 points) [from chapter 2 of https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf ]

What is the distance between two parallel hyperplanes $\{x \in \mathbb{R}^n \mid a^\top x = b_1\} \quad \text{and} \quad \{x \in \mathbb{R}^n \mid a^\top x = b_2\}$?

**Problem 1 Ans**

We let two parallel hyperplanes be:

$$\{x \in \mathbb{R}^n \mid a^\top x = b_1\} \quad \text{and} \quad \{x \in \mathbb{R}^n \mid a^\top x = b_2\}$$

The distance between hyperplanes that are parallel be calculated by $x_0$ projecting any point found in the first hyperplane $H$ onto the other hyperplane $H'$, given that $a^\top x_0 = b$. The hyperplane $H$ can be expressed as $H = \{x \in \mathbb{R}^n \mid a^\top (x - x_0) = 0\}$. The right hand term expressed as $\{x \in \mathbb{R}^n \mid a^\top (x - x_0) = 0\} = x_0 + a$, where $a^\top$ is the orthoganol of $a$

two parallel hyperplanes with the same the normal can be rewritten as

$H_1 = \{x \in \mathbb{R}^n \mid a^\top x = b_1\} = x_1 + a^\top$ and $H_2 = \{x \in \mathbb{R}^n \mid a^\top x = b_1\} = x_2 + a^\top$

Choosing any point $x_1 \in \mathbb{H}_1$, the displacement vector to $x_1 \in \mathbb{H}_1$ is just x_2-x_1.

The minimum distance between these two points lies on the normal vector $a4 such that

$$project_a(x_2 - x_1) = \frac{a^\top (x_2 - x_1)}{\|a\|^2} a$$

$a^\top x_1 = b_1$ and $a^\top x_2 = b_2$, so $a^\top (x_2 - x_1) = b_2 - b_1$. Making this substitution, the distance between hyperplanes that are parallel is

$$\frac{a^\top (x_2 - x_1)}{\|a\|^2} a = \frac{|a^\top (x_2 - x_1)|}{\|a\|} = \frac{|b_2 - b_1|}{\|a\|}$$

**Problem 2** (2 point) [from chapter 2 of https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf ]

When does one halfspace contain another? \ Give conditions under which $\{x \mid a^\top x \leq b\} \subseteq \{x \mid \tilde{a}^\top x \leq \tilde{b}\}$ (where $a \neq 0$, $\tilde{a} \neq 0$). Also find the conditions under which the two halfspaces are equal.

**Problem 2 ANS**

**(Part 1)** Conditions for when one halfspace contains

$$\{x \mid a^\top x \leq b\} \subseteq \{x \mid \tilde{a}^\top x \leq \tilde{b}\}$$

where $a \neq 0$, $\tilde{a} \neq 0$, the containment implies $H \subseteq \tilde{H}$ implies that every $x$ that satisfies $a^\top x \leq b$ must also satisfy $\tilde{a}^\top x \leq \tilde{b}$.

The containment only holds true if there exists a scalar $\lambda \geq 0$, under the conditions

$$\tilde{a} = \lambda a \quad \text{and} \quad \tilde{b} \geq \lambda b$$

- $\tilde{a}$ needs to be a positive multiple of $a$ so the hyperplanes are parallel with the same normal direction.
- When $\tilde{a} = \lambda a$ and $\lambda$ is positive, then the maximum of $\tilde{a}^\top x$ over $a^\top x \leq b$ is $\lambda b \leq \tilde{b}$, then containment holds if and only if $\tilde{b} \geq \lambda b$.

**(Part 2)**

Equality of halfspaces $H = \tilde{H}$ implies that there is containment in both directions,

$$H \subseteq \tilde{H} \quad \text{and} \quad \tilde{H} \subseteq H.$$

For two hyperplanes to be similar, the hyperplane equations must be proportional so the conditions

$$\tilde{a} = \lambda a \quad \text{and} \quad \tilde{b} = \lambda b$$

must be met for some $\lambda \geq 0$.

- $\tilde{a} = \lambda a$ establishes that the two hyperplanes have normal vectors that are parallel.
- The boundary equations are $a^\top x = b$ and $\tilde{a}^\top x = \tilde{b}$, making a substitution $(\tilde{a} = \lambda a)$ gives the equation $\lambda a^\top x = \tilde{b}$.
- Since the same boundary hyperplane is needed for equality, the equations must match $\tilde{a}^\top x = \lambda a^\top x$, thus giving $\tilde{b} = \lambda b$.

**Problem 3** (2 point) [from chapter 2 of https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf ]

Voronoi description of halfspace. Let $a$ and $b$ be distinct points in $\mathbb{R}^n$. Show that the set of all points that are closer (in Euclidean norm) to $a$ than $b$, i.e., $\{x \mid \|x - a\|_2 \leq \|x - b\|_2\}$, is a halfspace. Describe it explicitly as an inequality of the form $c^\top x \leq d$.

**Problem 3 ANS**

Define the set of points closer to $a$ than $b$ as

$$S = \{x \mid \|x - a\|_2 \leq \|x - b\|_2\}.$$

Since a Euclidean norm is nonnegative, squaring the term will maintain the inequality.

$$\|x - a\|_2^2 \leq \|x - b\|_2^2$$

Each term can be expanded by using $\|u\|_2^2 = u^\top u$ to give,

$$(x-a)^\top(x-a) \leq (x-b)^\top(x-b).$$

This can be further expanded as

$$x^\top x - 2a^\top x + a^\top a \leq x^\top x - 2b^\top x + b^\top b.$$

Rearranging the terms gives

$$2b^\top x - 2a^\top x + x^\top x - x^\top x \leq b^\top b - a^\top a,$$
$$= 2(b-a)^\top x \leq b^\top b - a^\top a.$$

Let $c = 2(b-a)$ and $d = b^\top b - a^\top a$ and making the substitution into the inequality gives

$$c^\top x \leq d.$$

The set is a halfspace with the form

$$\{x \in \mathbb{R}^n \mid c^\top x \leq d\}.$$

The points that are equidistant to $a$ and $b$ are given by a hyperplane in which the normal is in the direction of $b - a$.

**Problem 4** (4 points) [ use `nltk` Python library and https://www.nltk.org/book/ ]

- Use the state of the union addresses to create a term document matrix. Perform lower casing and stemming and, remove the non-alphabets and stop words.
- Let $p$ and $n$ stand for the number of words and documents, respectively. After preprocessing as stated above, what is $p, n$?
- Create the TF-IDF matrix and do everything below in terms of it.
- Compute the $\ell_2$ distance between the documents.
- Apply K-means for $K = 1, \cdots, 20$. Plot within-cluster sum of squares (WCSS) as a function of $K$. Use this graph to decide on the appropriate number of clusters.
- Print the name of the presidents, along with the year of the state of the union addresses that fall within the same cluster. Make it presentable and concise.
- Create a heatmap of the dissimilarity by putting documents that are in the same cluster next to each other.

In [9]:
```
import nltk
from nltk.corpus import state_union, stopwords
from nltk.stem import PorterStemmer
import re
import numpy as np
```

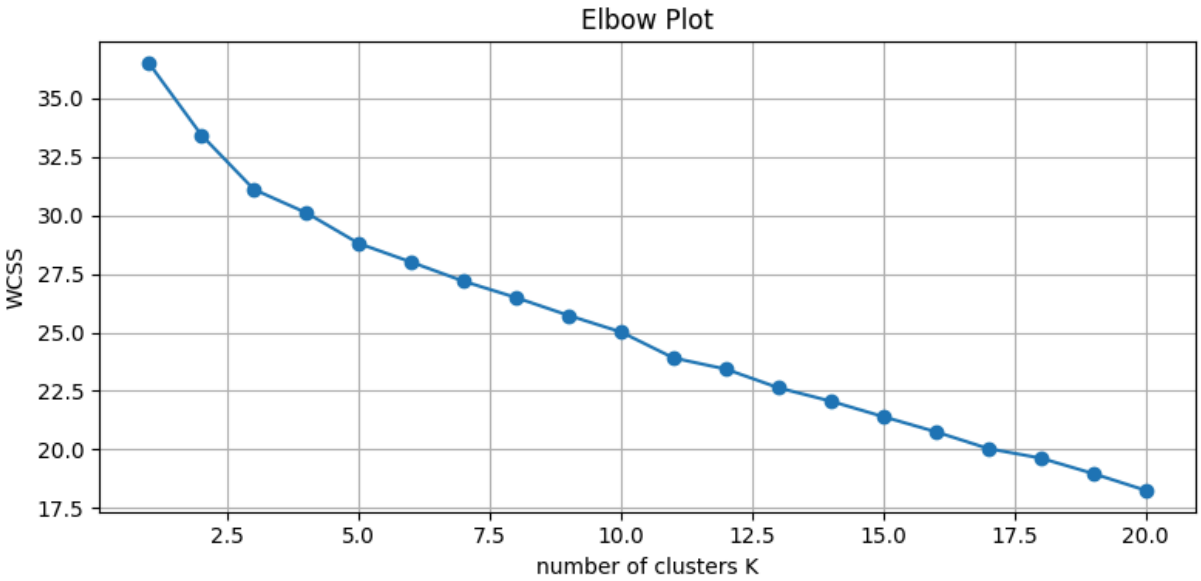
```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import pairwise_distances
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# --- Load files ---
fileids = state_union.fileids()
documents = [state_union.raw(fid) for fid in fileids]
n = len(documents)
print(f"Number of documents (n) = {n}")

# --- Preprocessing tokenizer ---
stemmer = PorterStemmer()
stop_words = set(stopwords.words('english'))

def tokenize_and_stem(text):
    text = text.lower()
    # keep alphabetic sequences only
    tokens = re.findall(r"[a-z]+", text)
    # remove stopwords
    tokens = [t for t in tokens if t not in stop_words]
    # stem
    stems = [stemmer.stem(t) for t in tokens]
    return stems

# --- TF-IDF matrix ---
vectorizer = TfidfVectorizer(tokenizer=tokenize_and_stem, lowercase=False)  # lower
tfidf = vectorizer.fit_transform(documents)
p = len(vectorizer.get_feature_names_out())
print(f"Vocabulary size after preprocessing (p) = {p}")

# --- Pairwise L2 distances between documents (Euclidean on TF-IDF rows) ---
tfidf_array = tfidf.toarray()
dists = pairwise_distances(tfidf_array, metric='euclidean')

# Show a small sample of pairwise distances
print("Pairwise L2 distance sample (first 5 docs):")
print(pd.DataFrame(dists[:5, :5], columns=fileids[:5], index=fileids[:5]))

# --- KMeans for K = 1..20, collect WCSS ---
wcss = []
k_range = list(range(1, 21))
for k in k_range:
    km = KMeans(n_clusters=k, random_state=42, n_init=10)
    km.fit(tfidf)
    wcss.append(km.inertia_)
print("Done computing WCSS for K=1..20")

# --- Plot the elbow (WCSS vs K) ---
plt.figure(figsize=(8,4))
plt.plot(k_range, wcss, marker='o')
plt.xlabel('number of clusters K')
plt.ylabel('WCSS')
plt.title('Elbow Plot')
plt.grid(True)
```

```
plt.tight_layout()
plt.show()

# --- Automatic elbow suggestion (simple relative-drop heuristic) ---
drops = np.diff(wcss)  # negative numbers (decrease)
relative_drops = -drops / wcss[:-1]
suggested_k = int(np.argmax(relative_drops) + 2)  # +2 because diff index correspon
print(f"Suggested K by relative-drop heuristic: {suggested_k}")

# You may also inspect the elbow plot visually and decide a K.
K = suggested_k

# --- Fit final KMeans with chosen K and get assignments ---
kmeans_final = KMeans(n_clusters=K, random_state=42, n_init=20)
clusters = kmeans_final.fit_predict(tfidf)

# --- Build metadata: president and year from fileid like '1946-Truman.txt' ---
meta = []
for fid, cluster in zip(fileids, clusters):
    label = fid.replace('.txt','')
    parts = label.split('-')
    year = parts[0]
    pres = '-'.join(parts[1:]) if len(parts) > 1 else ''
    meta.append({"fileid": fid, "year": year, "president": pres, "cluster": int(clu

meta_df = pd.DataFrame(meta)
meta_df_sorted = meta_df.sort_values(['cluster','year']).reset_index(drop=True)

# --- Presentable concise printout of clusters ---
print("\n=== Clusters (President - Year) grouped by cluster ===")
for c in sorted(meta_df_sorted['cluster'].unique()):
    group = meta_df_sorted[meta_df_sorted['cluster']==c][['year','president']]
    lines = [f"{r['year']}: {r['president']}" for r in group.to_dict('records')]
    print(f"\nCluster {c} ({len(lines)} docs):")
    print("  " + "; ".join(lines))

# --- Create heatmap of dissimilarity with docs in same cluster adjacent ---
order = np.argsort(clusters)            # sort documents by cluster id
dists_ordered = dists[np.ix_(order, order)]
labels_ordered = [fileids[i] for i in order]

plt.figure(figsize=(10,8))
plt.imshow(dists_ordered, aspect='auto')
plt.colorbar()
plt.title(f'Heatmap of L2 distances (docs ordered by cluster)')
plt.xticks(ticks=np.arange(len(labels_ordered)), labels=labels_ordered, rotation=90
plt.yticks(ticks=np.arange(len(labels_ordered)), labels=labels_ordered, fontsize=6)
plt.tight_layout()
plt.show()
```

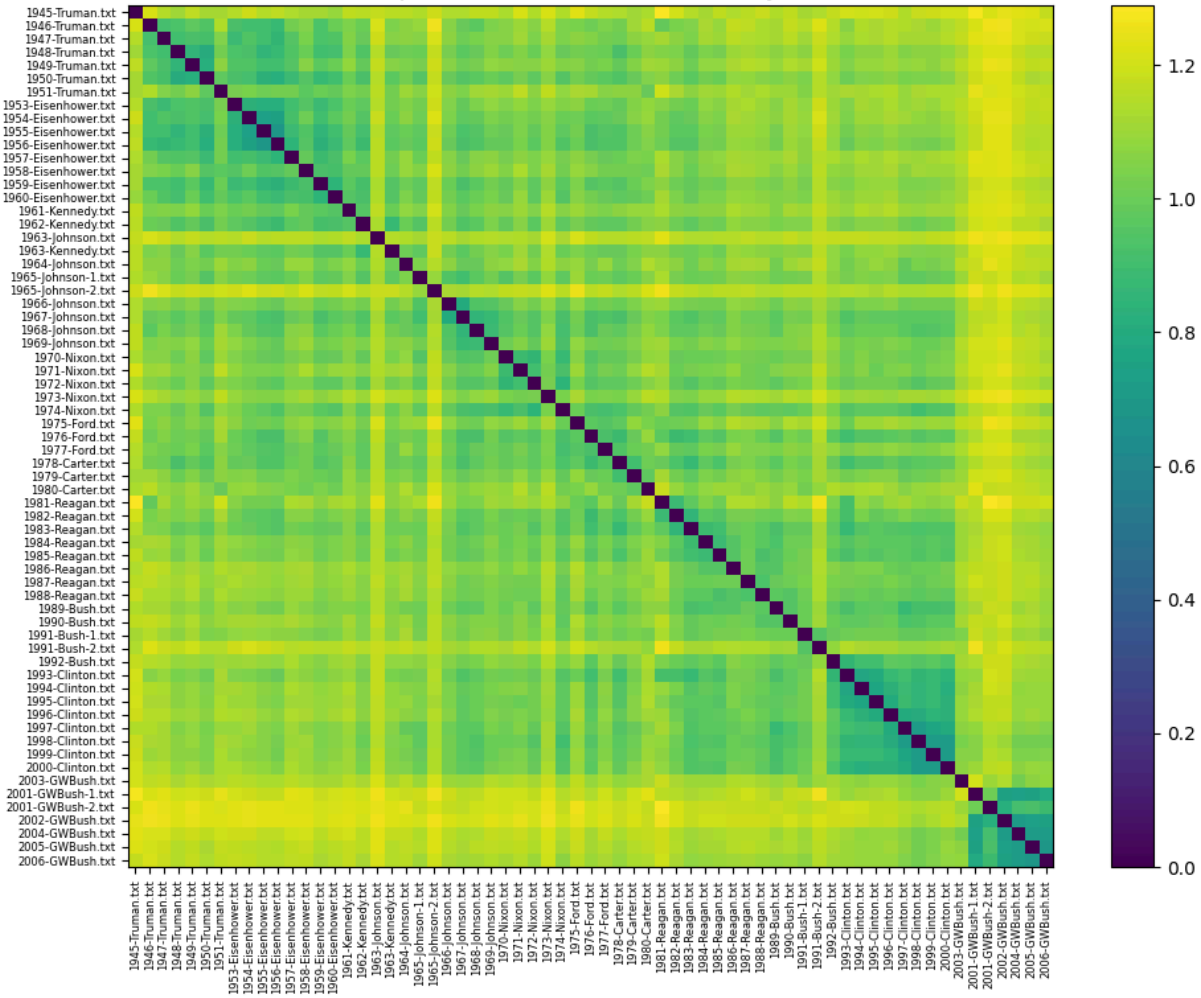
```
Number of documents (n) = 65
```

```
c:\Users\saad_\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\fea
ture_extraction\text.py:517: UserWarning: The parameter 'token_pattern' will not be
used since 'tokenizer' is not None'
  warnings.warn(
```

```
Vocabulary size after preprocessing (p) = 7280
Pairwise L2 distance sample (first 5 docs):
                 1945-Truman.txt  1946-Truman.txt  1947-Truman.txt  \
1945-Truman.txt         0.000000         1.216738         1.147651
1946-Truman.txt         1.216738         0.000000         0.914376
1947-Truman.txt         1.147651         0.914376         0.000000
1948-Truman.txt         1.109533         0.934377         0.896245
1949-Truman.txt         1.164174         0.975824         0.927224

                 1948-Truman.txt  1949-Truman.txt
1945-Truman.txt         1.109533         1.164174
1946-Truman.txt         0.934377         0.975824
1947-Truman.txt         0.896245         0.927224
1948-Truman.txt         0.000000         0.853042
1949-Truman.txt         0.853042         0.000000
Done computing WCSS for K=1..20
```




```
Suggested K by relative-drop heuristic: 2

=== Clusters (President - Year) grouped by cluster ===

Cluster 0 (59 docs):
  1945: Truman; 1946: Truman; 1947: Truman; 1948: Truman; 1949: Truman; 1950: Truma
n; 1951: Truman; 1953: Eisenhower; 1954: Eisenhower; 1955: Eisenhower; 1956: Eisenho
wer; 1957: Eisenhower; 1958: Eisenhower; 1959: Eisenhower; 1960: Eisenhower; 1961: K
ennedy; 1962: Kennedy; 1963: Johnson; 1963: Kennedy; 1964: Johnson; 1965: Johnson-1;
1965: Johnson-2; 1966: Johnson; 1967: Johnson; 1968: Johnson; 1969: Johnson; 1970: N
ixon; 1971: Nixon; 1972: Nixon; 1973: Nixon; 1974: Nixon; 1975: Ford; 1976: Ford; 19
77: Ford; 1978: Carter; 1979: Carter; 1980: Carter; 1981: Reagan; 1982: Reagan; 198
3: Reagan; 1984: Reagan; 1985: Reagan; 1986: Reagan; 1987: Reagan; 1988: Reagan; 198
9: Bush; 1990: Bush; 1991: Bush-1; 1991: Bush-2; 1992: Bush; 1993: Clinton; 1994: Cl
inton; 1995: Clinton; 1996: Clinton; 1997: Clinton; 1998: Clinton; 1999: Clinton; 20
00: Clinton; 2003: GWBush

Cluster 1 (6 docs):
  2001: GWBush-1; 2001: GWBush-2; 2002: GWBush; 2004: GWBush; 2005: GWBush; 2006: GW
Bush
```


Heatmap of L2 distances (docs ordered by cluster)
1945-Truman.txt
1946-Truman.txt
1947-Truman.txt
1948-Truman.txt
1949-Truman.txt
1950-Truman.txt
1951-Truman.txt
1953-Eisenhower.txt
1954-Eisenhower.txt
1955-Eisenhower.txt
1956-Eisenhower.txt
1957-Eisenhower.txt
1958-Eisenhower.txt
1959-Eisenhower.txt
1960-Eisenhower.txt
1961-Kennedy.txt
1962-Kennedy.txt
1963-Johnson.txt
1963-Kennedy.txt
1964-Johnson.txt
1965-Johnson-1.txt
1965-Johnson-2.txt
1966-Johnson.txt
1967-Johnson.txt
1968-Johnson.txt
1969-Johnson.txt
1970-Nixon.txt
1971-Nixon.txt
1972-Nixon.txt
1973-Nixon.txt
1974-Nixon.txt
1975-Ford.txt
1976-Ford.txt
1977-Ford.txt
1978-Carter.txt
1979-Carter.txt
1980-Carter.txt
1981-Reagan.txt
1982-Reagan.txt
1983-Reagan.txt
1984-Reagan.txt
1985-Reagan.txt
1986-Reagan.txt
1987-Reagan.txt
1988-Reagan.txt
1989-Bush.txt
1990-Bush.txt
1991-Bush-1.txt
1991-Bush-2.txt
1992-Bush.txt
1993-Clinton.txt
1994-Clinton.txt
1995-Clinton.txt
1996-Clinton.txt
1997-Clinton.txt
1998-Clinton.txt
1999-Clinton.txt
2000-Clinton.txt
2003-GWBush.txt
2001-GWBush-1.txt
2001-GWBush-2.txt
2002-GWBush.txt
2004-GWBush.txt
2005-GWBush.txt
2006-GWBush.txt
1.2
1.0
0.8
0.6
0.4
0.2
0.0