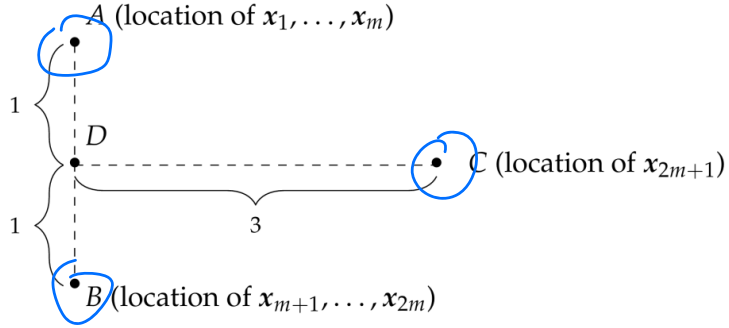
HW3 Review

CSCI 567 @ Fall 2025

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.1 First, consider running the greedy initialization method, where the first center is selected uniformly at random from the training set, and then the point farthest away from the first center is selected as the second center (tie broken arbitrarily).

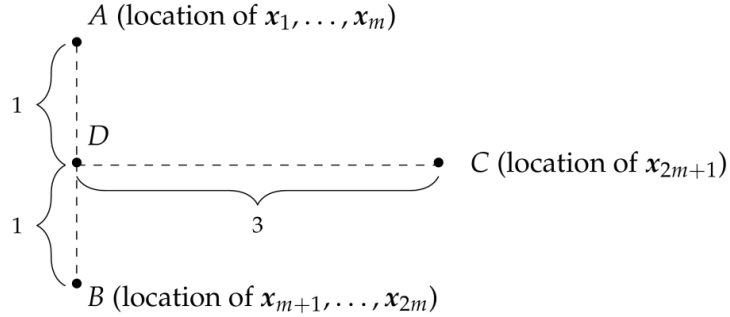
a) What are the possible outcomes of greedy initialization? Explain briefly.

b) What are the final outputs of K-means under this initialization strategy? Describe the location of centers and the assignments of each point, and explain briefly.

c) What is the final value for the K-means objective in this case?

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.1 First, consider running the greedy initialization method, where the first center is selected uniformly at random from the training set, and then the point farthest away from the first center is selected as the second center (tie broken arbitrarily).

a) What are the possible outcomes of greedy initialization? Explain briefly.

{A, C} or {B, C}.

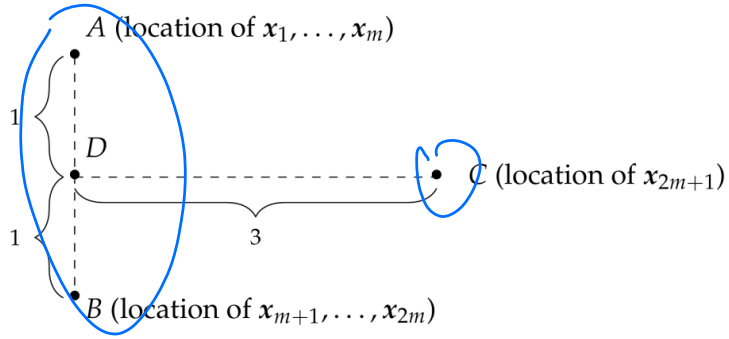
Reason: If $\mu_1 = A$ or B, then $\mu_2 = x_{2m+1} = C$ since x_{2m+1} is the (only) farthest point from both A and B. If $\mu_1 = C$, then $\mu_2 \in \{A, B\}$.

b) What are the final outputs of K-means under this initialization strategy? Describe the location of centers and the assignments of each point, and explain briefly.

c) What is the final value for the K-means objective in this case?

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.1 First, consider running the greedy initialization method, where the first center is selected uniformly at random from the training set, and then the point farthest away from the first center is selected as the second center (tie broken arbitrarily).

a) What are the possible outcomes of greedy initialization? Explain briefly.

{A, C} or {B, C}.

Reason: If $\mu_1 = A$ or B, then $\mu_2 = x_{2m+1} = C$ since x_{2m+1} is the (only) farthest point from both A and B. If $\mu_1 = C$, then $\mu_2 \in \{A, B\}$.

b) What are the final outputs of K-means under this initialization strategy? Describe the location of centers and the assignments of each point, and explain briefly.

Final centers: {C, D}

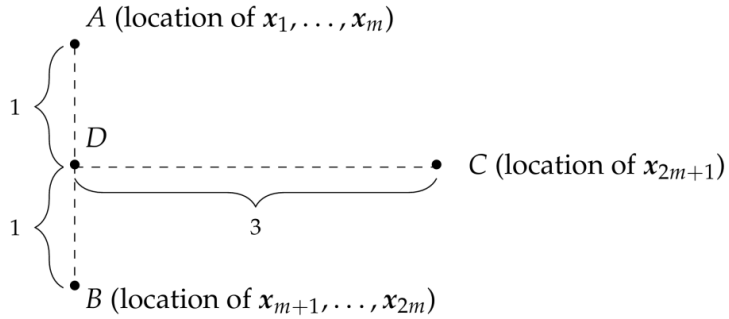
Assignments: Only x_{2m+1} assigned to C; all others assigned to D

Reason:

- Suppose the init. centers are $\{A, C\}$. At the first iter., all points at A and B will be assigned to the same cluster, while x_{2m+1} will be assigned to the other.
- By averaging the points in each cluster, we obtain new centers D and C.
- By symmetry, the same holds when the init. centers are {B, C}

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.1 First, consider running the greedy initialization method, where the first center is selected uniformly at random from the training set, and then the point farthest away from the first center is selected as the second center (tie broken arbitrarily).

a) What are the possible outcomes of greedy initialization? Explain briefly.

{A, C} or {B, C}. Reason: (previous page)

b) What are the final outputs of K-means under this initialization strategy? Describe the location of centers and the assignments of each point, and explain briefly.

Final centers: {C, D}

Assignments: Only x_{2m+1} assigned to C; all others assigned to D

Reason: (previous page)

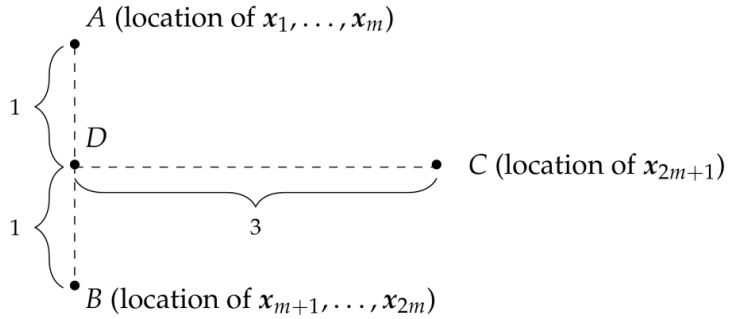
c) What is the final value for "sum of squared distances" objective in this case?

$$\|x_{2m+1} - C\|_2^2 + \sum_{i=1}^{2m} \|x_i - D\|_2^2 = 2m$$

$= 0$ $\quad$ 1

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.2 Next consider running K-means++

a) What is the probability for each pair of the initial centers $\{A, B\}, \{A, C\}, \{B, C\}$?

Slightly abusing notation: let A_i, B_i, C_i denote the events that the ith center is initialized to A, B, or C (respectively). Notice that

$$\mathbb{P}(A_1) = \frac{m}{2m+1} = \mathbb{P}(B_1), \qquad \mathbb{P}(C_1) = \frac{1}{2m+1}$$
$$\|B - A\|_2^2 = 4, \qquad \|C - A\|_2^2 = 10 = \|C - B\|_2^2$$

Therefore, we have

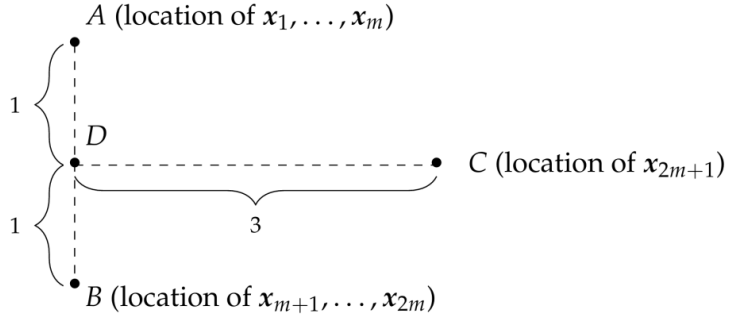
$$\mathbb{P}(A_2|B_1) = \mathbb{P}(B_2|A_1) = \frac{4m}{4m+10} = \frac{2m}{2m+5}$$
$$\Rightarrow \mathbb{P}(\text{init. to } \{\text{A, B}\}) = \mathbb{P}(A_1 B_2) + \mathbb{P}(A_2 B_1) = 2 \cdot \frac{m}{2m+1} \cdot \frac{2m}{2m+5}$$

Moreover,

$$\mathbb{P}(C_2|A_1) = \mathbb{P}(C_2|B_1) = \frac{10}{4m+10}, \qquad \mathbb{P}(A_2|C_1) = \mathbb{P}(B_2|C_1) = \frac{1}{2}$$
$$\mathbb{P}(\text{init. to } \{\text{A, C}\}) = \mathbb{P}(\text{init. to } \{\text{B, C}\}) = \frac{1}{2} \cdot \frac{1}{2m+1} + \frac{m}{2m+1} \cdot \frac{5}{2m+5}$$

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:

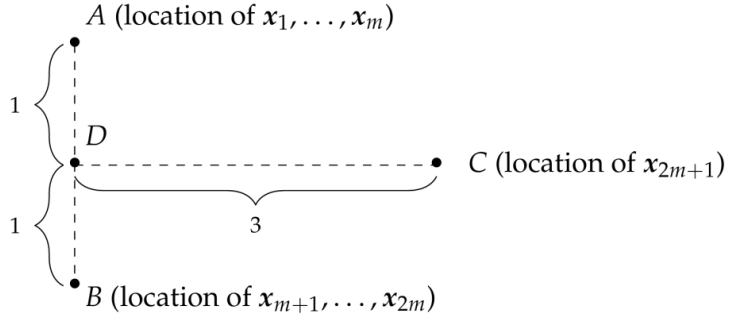


1.2. b) When the initial centers happen to be at A and B, the outlier x_{2m+1} will eventually be clustered together with either the points at A or the points at B (depending on how you break tie). For simplicity, we assume that m is very large such that the final centers of the two cluster can be approximately treated as A and B. Under this assumption, calculate the final K-means objective value in this case. (1 point)

Clearly, only the distance between x_{2m+1} and its center is nonzero. The square of this distance is exactly 10, which is also the final K-means objective value.

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.2. C)

Combining your solutions from all previous questions, write down the expected final K-means objective value when running K-means++, and calculate its limit when m goes to infinity. (3 points)

Based on previous questions, the expected final K-means objective value is

$\mathbb{P}(\text{init. to } \{A, B\}) \times \mathbb{E}[\text{Value} | \text{final centers} = \{A, B\}]$

$$\left(\frac{2m}{2m+1} \times \frac{2m}{2m+5}\right) \times 10 + \left(\frac{1}{2m+1} + \frac{2m}{2m+1} \times \frac{5}{2m+5}\right) \times 2m.$$

$\mathbb{P}(\text{init. to } \{A, C\} \text{ or } \{B, C\}) \times \mathbb{E}[\text{Value} | \text{final centers} = \{D, C\}]$

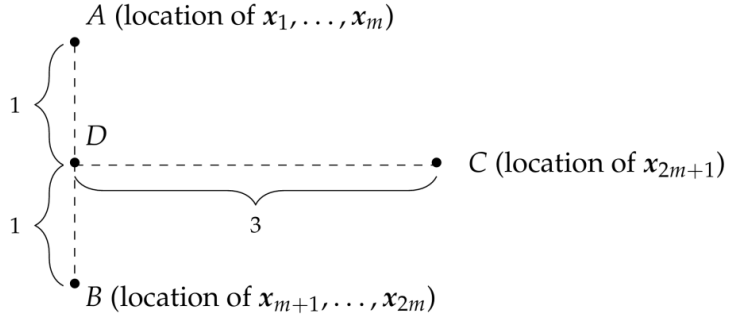
Writing this as

$$\left(\frac{2}{2+\frac{1}{m}} \times \frac{2}{2+\frac{5}{m}}\right) \times 10 + \left(\frac{2}{2+\frac{1}{m}} + \frac{2}{2+\frac{1}{m}} \times \frac{10}{2+\frac{5}{m}}\right),$$

we see that the limit when $m \to \infty$ is clearly $10 + (1 + 5) = 16$.

Problem 1. Clustering

Consider running K-means with $K = 2$ on the following dataset:



1.3 Based on Q1.1 and Q1.2, explain briefly why K-means++ is better than greedy initialization in this dataset.

For greedy, the objective value $= 2m \to \infty$ as $m \to \infty$
For K-means++, the objective value is always bounded.
$\Rightarrow$ K-means++ is better

Problem 2. EM

Known:

- data $x_1, \dots, x_N \in \mathbb{R}^D$
- noise scale $\sigma > 0$

Unknown:

- Latent $z_1, \dots, z_N \in \mathbb{R}$
- Parameter $v \in \mathbb{R}^d$

Model:

$p(z) \propto \exp\left(-\frac{1}{2}z^2\right)$

$p(x \mid z; v) \propto \exp\left(-\frac{1}{2\sigma^2}\|x - zv\|_2^2\right)$

$$b^2 = \frac{\sigma^2}{\sigma^2 + \|v\|_2^2} = \left(1 + \frac{\|v\|_2^2}{\sigma^2}\right)^{-1}$$

$$a_n = \frac{x_n^T v}{\sigma^2 + \|v\|_2^2} = \frac{x_n^T v}{\sigma^2} \cdot b^2$$

$$z^2 - 2a_n z = (z - a_n)^2 - a_n^2$$

2.1 E-step 1: fixing the parameter v, prove that the posterior distribution $q_n(z) \triangleq p(z_n = z \mid x_n; v)$ is also a one-dimensional Gaussian distribution. More specifically, show that

$$q_n(z) \propto \exp\left(-\frac{1}{2b^2}(z - a_n)^2\right),$$

where the mean $a_n = \frac{x_n^\top v}{\sigma^2 + \|v\|_2^2}$ and the variance $b^2 = \frac{\sigma^2}{\sigma^2 + \|v\|_2^2}$. (6 points)

$$\begin{aligned}
p(z_n = z \mid x_n; v) &\propto p(z_n = z, x_n; v) \\
&= p(z)p(x_n \mid z; v) \quad \text{by Chain rule.} \\
&\propto \exp\left(-\frac{1}{2}z^2\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\|x_n - zv\|_2^2\right) \\
&= \exp\left(-\frac{1}{2}\left(z^2 + \frac{\|x_n\|_2^2}{\sigma^2} + \frac{\|v\|_2^2}{\sigma^2}z^2 - \frac{2x_n^T v}{\sigma^2}z\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(1 + \frac{\|v\|_2^2}{\sigma^2}\right)z^2 + \frac{x_n^\top v}{\sigma^2}z\right) \\
&= \exp\left(-\frac{1}{2b^2}\left(z^2 - 2b^2 \cdot \frac{x_n^T v}{\sigma^2}z\right)\right) \\
&= \exp\left(-\frac{1}{2b^2}(z - a_n)^2 + \frac{a_n^2}{2b^2}\right) \\
&\propto \exp\left(-\frac{1}{2b^2}(z - a_n)^2\right)
\end{aligned}$$

$P(A|B) = \frac{P(AB)}{P(B)} \propto P(AB)$.

$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$

Problem 2. EM

Known:

- data $x_1, \ldots, x_N \in \mathbb{R}^D$
- noise scale $\sigma > 0$

Unknown:

- Latent $z_1, \ldots, z_N \in \mathbb{R}$
- Parameter $v \in \mathbb{R}^d$

Model:

$p(z) \propto \exp\left(-\frac{1}{2}z^2\right)$

$p(x \mid z; v) \propto \exp\left(-\frac{1}{2\sigma^2}\|x - zv\|_2^2\right)$

$q_n(z) = p(z_n = z \mid x_n; v)$
$\propto \exp\left(-\frac{1}{2b^2}(z - a_n)^2\right)$

$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

2nd moment

2.2 E-step 2: write down the expected complete log-likelihood $Q(v)$. Express it in terms of v, σ, x_n, a_n, and b^2 (for $n = 1, \ldots, N$), and feel free to drop any terms independent of v. You can solve this using conclusions from **E-step 1**, even if you have not solved it yet. (Note that in the lecture, we write Q in terms of the parameter, which is v here, and also its previous value; here, the previous value of v is already used in defining a_n and b^2, which is why you do not need to write Q using the previous value of v explicitly.) (5 points)

First, we have

by chain rule

$$Q(v) = \sum_{n=1}^{N} \mathbb{E}_{z_n \sim q_n}[\ln p(x_n, z_n; v)] = \sum_{n=1}^{N} \mathbb{E}_{z_n \sim q_n}[\ln p(z_n) + \ln p(x_n \mid z_n; v)]. \quad \text{(1 point)}$$

Plugging in the probabilistic model and dropping all terms independent of v, we have

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} \mathbb{E}_{z_n \sim q_n}[\|x_n - z_n v\|_2^2]. \quad \text{(1 point)}$$

Expanding the square and dropping the terms independent of v again gives:

dropped $\|x_n\|_2^2$.

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} \mathbb{E}_{z_n \sim q_n}\left[z_n^2 \|v\|_2^2 - 2 z_n x_n^\top v\right]. \quad \text{(1 point)}$$

Finally, we use the facts

$$\mathbb{E}_{z_n \sim q_n}[z_n] = a_n, \quad \mathbb{E}_{z_n \sim q_n}[z_n^2] = b^2 + (\mathbb{E}_{z_n \sim q_n}[z_n])^2 = b^2 + a_n^2, \quad \text{(1 point)}$$

to arrive at

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left((b^2 + a_n^2)\|v\|_2^2 - 2 a_n x_n^\top v\right) = -\frac{1}{2\sigma^2}\left(\left(\sum_{n=1}^{N}(b^2 + a_n^2)\right)\|v\|_2^2 - 2\left(\sum_{n=1}^{N} a_n x_n\right)^\top v\right). \quad \text{(1 point)}$$

Problem 2. EM

Known:

- data $x_1, \ldots, x_N \in \mathbb{R}^D$
- noise scale $\sigma > 0$

Unknown:

- Latent $z_1, \ldots, z_N \in \mathbb{R}$
- Parameter $v \in \mathbb{R}^d$

Model:

$p(z) \propto \exp\left(-\frac{1}{2}z^2\right)$

$p(x \mid z; v) \propto \exp\left(-\frac{1}{2\sigma^2}\|x - zv\|_2^2\right)$

$$Q(v) \propto -\frac{1}{2\sigma^2}\left(\left(\sum_{n=1}^{N}(b^2 + a_n^2)\right)\|v\|_2^2 - 2\left(\sum_{n=1}^{N} a_n x_n\right)^\top v\right)$$

2.3 M-step: Find the maximizer of $Q(v)$ by setting its gradient to $\mathbf{0}$. Express it in terms of x_n, a_n, and b^2 (for $n = 1, \ldots, N$). (3 points)

Simply setting the gradient to zero gives

$$2\left(\sum_{n=1}^{N}(b^2 + a_n^2)\right)v - 2\left(\sum_{n=1}^{N} a_n x_n\right) = \mathbf{0}.$$

Solving for v gives

$$v = \frac{\sum_{n=1}^{N} a_n x_n}{\sum_{n=1}^{N}(b^2 + a_n^2)}.$$

Rubrics: 2 points for the correct gradient and 1 point for the correct final answer.

Final note: in case you are wondering why this model is useful, it is in fact a greatly simplified version of something called "probabilistic PCA".

Problem 3. PCA

3.1 Specifically, suppose we have a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ with zero mean, and we would like to compress it into a one-dimensional dataset $c_1, \ldots, c_N \in \mathbb{R}$. To reconstruct the dataset (approximately), we also keep a direction vector $\mathbf{v} \in \mathbb{R}^D$ with unit norm (i.e. $\|\mathbf{v}\|_2 = 1$) so that the reconstructed dataset is $c_1\mathbf{v}, \ldots, c_N\mathbf{v} \in \mathbb{R}^D$.

The way we find $c_1, \ldots, c_N$ and $\mathbf{v}$ is to minimize the reconstruction error in terms of the squared L2 distance, that is, we solve

$$\arg\min_{c_1,\ldots,c_N,\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \|\mathbf{x}_n - c_n\mathbf{v}\|_2^2. \tag{1}$$

Prove that the solution of (1) is exactly the following

1) $c_n = \mathbf{x}_n^T\mathbf{v}$ for each $n = 1, \ldots, N$; (3 points)

2) $\mathbf{v}$ is the first principal component of the dataset, that is, $\arg\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^T \left(\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^T\right)\mathbf{v}$. (3 points)

Hint: first prove 1) by fixing $\mathbf{v}$, then prove 2) using the conclusion of 1).

If $F(x_1, \ldots, x_n) = \sum_i f_i(x_i)$, then $\min_{x_{1:n}} F = \sum_i \min f_i$

$$\min_{(x,y)\in X\times Y} f(x,y) = \min_x \left(\min_y f(x,y)\right)$$

For any fixed $\mathbf{v}$ (with unit norm), clearly we can optimize over each c_n independently: (1 point)

$$\begin{aligned}
\arg\min_{c_n} \|\mathbf{x}_n - c_n\mathbf{v}\|_2^2 &= \arg\min_{c_n} \left(c_n^2\|\mathbf{v}\|_2^2 - (2\mathbf{x}_n^T\mathbf{v})c_n + \|\mathbf{x}_n\|_2^2\right) \\
&= \arg\min_{c_n} \left(c_n^2 - (2\mathbf{x}_n^T\mathbf{v})c_n\right) && (\|\mathbf{v}\|_2^2 = 1) \\
&= \arg\min_{c_n} \left(c_n - \mathbf{x}_n^T\mathbf{v}\right)^2 && \text{(1 point)} \\
&= \mathbf{x}_n^T\mathbf{v}. && \text{(1 point)}
\end{aligned}$$

Next, plugging $c_n = \mathbf{x}_n^T\mathbf{v}$ back into the objective, we see that $\mathbf{v}$ is the solution of

$$\begin{aligned}
\arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \|\mathbf{x}_n - (\mathbf{x}_n^T\mathbf{v})\mathbf{v}\|_2^2 &= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \left(\|\mathbf{x}_n\|_2^2 - 2(\mathbf{x}_n^T\mathbf{v})^2 + (\mathbf{x}_n^T\mathbf{v})^2\|\mathbf{v}\|_2^2\right) && \text{(1 point)} \\
&= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \left(-(\mathbf{x}_n^T\mathbf{v})^2\right) && (\|\mathbf{x}_n\|_2^2 \text{ is irrelevant and } \|\mathbf{v}\|_2^2 = 1, \text{1 point}) \\
&= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} -\sum_{n=1}^{N} \mathbf{v}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{v} && (\text{scalar}) \quad (\mathbf{x}_n^T\mathbf{v})(\mathbf{x}_n^T\mathbf{v}) = (\mathbf{v}^T\mathbf{x}_n)(\mathbf{x}_n^T\mathbf{v}) \\
&= \arg\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^T\left(\sum_{n=1}^{N} \mathbf{x}_n\mathbf{x}_n^T\right)\mathbf{v}. && \text{(1 point)}
\end{aligned}$$

Problem 3. PCA

3.2 Next, you are asked to generalize the same idea to an arbitrary compression dimension $p < D$. Specifically, we would like to compress the same zero-mean dataset into a p-dimensional dataset $\mathbf{c}_1, \ldots, \mathbf{c}_N \in \mathbb{R}^p$. To reconstruct the dataset (approximately), we also keep p orthogonal direction vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^D$ with unit norm. For notational convenience, we stack these vectors together as a matrix $\mathbf{V} \in \mathbb{R}^{D \times p}$ whose j-th column is $\mathbf{v}_j$.

1) Write down the reconstructed dataset using $\mathbf{c}_1, \ldots, \mathbf{c}_N$ and $\mathbf{V}$ (note: this is a set of points in $\mathbb{R}^D$). Then write down the analogue of (1), that is, the optimization problem (with variables $\mathbf{c}_1, \ldots, \mathbf{c}_N$ and $\mathbf{V}$) that minimizes the reconstruction error in terms of the squared L2 distance. Make sure to include the correct constraints in this optimization problem. No reasoning needed. (3 points)

The reconstructed dataset is $\mathbf{V}\mathbf{c}_1, \ldots, \mathbf{V}\mathbf{c}_N \in \mathbb{R}^D$. Thus the optimization problem is

Reconstruct $\hat{x}_n = c_{n,1} v_1 + \cdots + c_{n,p} v_p = V c_n$

$$\underset{\substack{\mathbf{c}_1,\ldots,\mathbf{c}_N \in \mathbb{R}^p \\ \mathbf{V} \in \mathbb{R}^{D \times p}:\ \mathbf{V}^\top \mathbf{V} = \mathbf{I}}}{\arg\min} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{V}\mathbf{c}_n\|_2^2,$$

where $\mathbf{I}$ is the $p \times p$ identity matrix.

Problem 3. PCA

3.2. 2) Find the optimal solution of $\mathbf{c}_1, \ldots, \mathbf{c}_N$ while fixing $\mathbf{V}$.

Similarly, the optimization can be done independently for each $\mathbf{c}_n$:

$$\begin{aligned}\arg\min_{\mathbf{c}_n} \|\mathbf{x}_n - \mathbf{V}\mathbf{c}_n\|_2^2 &= \arg\min_{\mathbf{c}_n} \left(\|\mathbf{x}_n\|_2^2 - 2\mathbf{x}_n^\top \mathbf{V}\mathbf{c}_n + \mathbf{c}_n^\top \mathbf{V}^\top \mathbf{V}\mathbf{c}_n \right) \\ &= \arg\min_{\mathbf{c}_n} \left(\mathbf{c}_n^\top \mathbf{c}_n - 2\mathbf{x}_n^\top \mathbf{V}\mathbf{c}_n \right).\end{aligned}$$

For the last step, simply setting the gradient $2\mathbf{c}_n - 2\mathbf{V}^\top \mathbf{x}_n$ to $\mathbf{0}$ gives the solution $\mathbf{c}_n = \mathbf{V}^\top \mathbf{x}_n$.

Rubrics: The breakdown of points is only for reference.

Problem 3. PCA

3.2. 3) Plug the solution of the previous question into the optimization problem and conclude that the optimal solution of $\mathbf{V}$ is (5 points)

$$\arg\max_{\mathbf{V}\in\mathbb{R}^{D\times p}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{j=1}^{p}\left(\mathbf{v}_j^\top\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top\right)\mathbf{v}_j\right).$$

$V = [v_1 \ \dots \ v_p]$
Col j is $v_j \in \mathbb{R}^{D\times 1}$

(This means that $\mathbf{v}_1,\dots,\mathbf{v}_p$ are exactly the top p eigenvectors of $\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top$, that is, the top p principal components. You do not need to prove this fact though.)

After plugging $\mathbf{c}_n = \mathbf{V}^\top\mathbf{x}_n$, the problem becomes

$$\begin{aligned}
\arg\min_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\|\mathbf{x}_n-\mathbf{V}\mathbf{V}^\top\mathbf{x}_n\|_2^2 &= \arg\min_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\left(\|\mathbf{x}_n\|_2^2-2\mathbf{x}_n^\top\mathbf{V}\mathbf{V}^\top\mathbf{x}_n+\mathbf{x}_n^\top\mathbf{V}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{x}_n\right) && \text{(1 point)}\\
&= \arg\min_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\left(-2\mathbf{x}_n^\top\mathbf{V}\mathbf{V}^\top\mathbf{x}_n+\mathbf{x}_n^\top\mathbf{V}\mathbf{V}^\top\mathbf{x}_n\right)\\
&&& (\|\mathbf{x}_n\|_2^2 \text{ is irrelevant and } \mathbf{V}^\top\mathbf{V}=\mathbf{I}\text{, 1 point})\\
&= \arg\max_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\left(\mathbf{x}_n^\top\mathbf{V}\mathbf{V}^\top\mathbf{x}_n\right) && \text{(1 point)}\\
&= \arg\max_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\left(\mathbf{x}_n^\top\left(\sum_{j=1}^{p}\mathbf{v}_j\mathbf{v}_j^\top\right)\mathbf{x}_n\right) && \text{(1 point)}\\
&= \arg\max_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\sum_{j=1}^{p}\left(\mathbf{x}_n^\top\mathbf{v}_j\right)\left(\mathbf{v}_j^\top\mathbf{x}_n\right)\\
&= \arg\max_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^{N}\sum_{j=1}^{p}\left(\mathbf{v}_j^\top\mathbf{x}_n\right)\left(\mathbf{x}_n^\top\mathbf{v}_j\right)\\
&= \arg\max_{\mathbf{V}:\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{j=1}^{p}\left(\mathbf{v}_j^\top\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top\right)\mathbf{v}_j\right). && \text{(1 point)}
\end{aligned}$$

$V = [v_1 \ \dots \ v_p]$

$VV^\top = \sum_{j=1}^{p} v_j v_j^\top$

$[v_1 \ \dots \ v_p]\begin{bmatrix} v_1^\top \\ \vdots \\ v_p^\top \end{bmatrix} = \sum_{j=1}^{p} v_j v_j^\top$

(block matrix multip.)