Andrew Li-Yang Liu
Nakul Verma
Unsupervised Learning
October 8, 2022

UML Homework 1

Problem 1: Summary of reading "Clustering large graphs via the singular value decomposition"

Goal and significance:

Discrete clustering problems (DCP) in Euclidean space are in general NP-hard, so this paper proposes a continuous relaxation of the problem (CCP) that can be solved efficiently using SVD. They call this continuous relaxation "generalized clustering", since each data point can now have partial/soft-assignment to each cluster. The paper then proposes an efficient randomized algorithm for approximating the SVD. These results enable efficient (approximate) clustering of the rows of large matrices, and can then be applied to clustering large graphs.

Summary of algorithms:

Continuous relaxation (CCP) and solving via SVD:

More concretely, the paper transforms the discrete k-cluster problem into a dimensionality reduction of the data points onto a k-dimensional subspace. Whereas, given m points $A = \{A_1, \dots, A_m\}$, the goal of DCP is to find cluster centers $B = \{B_1, \dots, B_k\}$ that minimizes the following cost:

$$f_A(B) = \sum_{i=1}^{m} dist^2(A_i, B)$$

CCP finds the k-dimensional subspace V that minimizes:

$$g_A(V) = \sum_{i=1}^{m} dist^2(A_i, V)$$

Note that if $\bar{A}_1, \dots, \bar{A}_m$ are the orthogonal projections of $A_1, \dots, A_m$ onto k-dimensional subspace V, then:

$$g_A(V) = \sum_{i=1}^{m} dist^2(A_i, V) = \sum_{i=1}^{m} dist^2(A_i, \bar{A}_i) = \left||A - \bar{A}|\right|_F^2$$

Moreover, $\bar{A}$ has rank k, so this minimization problem is solved using the truncated SVD for $\bar{A}$ (due to the Eckart-Young theorem).

Now that we have $\bar{A}$, we solve the DCP using the m points $\bar{A}_1, \ldots, \bar{A}_m \in R^k$ to obtain $\bar{B}$, the k-centroids in R^k. Return $\bar{B}$ as the approximation for B.

Generalized clustering for large graphs:
The authors propose using SVD to cluster large graphs. The goal is to create a process that removes the maximum weight cluster from a graph.

Consider the cluster $x \in \mathbb{R}^m$, where $x(u)$ is the intensity with which u belongs to cluster x and $\|x\| = 1$. We can then consider the quantity: $x^T A$. Note that $(x^T A)_i$ is the frequency of node I occurring in neighborhood of cluster x, so high $|(x^T A)_i|$ means high reinforcement. $\|x^T A\|_2^2$ is thus a measure of importance of cluster x.

Requiring two clusters u and v to be different is not enough, since they may be arbitrarily close to each other. Hence, we must require cluster vectors to be orthogonal. Note that requiring orthogonality is a relaxation of the disjointness requirement for clusters.

The optimal weight clustering of A is thus a set of orthonormal vectors $x^{(1)}, \ldots, x^{(k)}$ such that the $x^{(i)}$'s are maximum weight clusters of A. This amounts to selecting the largest singular vectors of A.

Removing the first k clusters is thus expressed as:

$$R^{(k)} = A - \sum_{t=1}^{k} x^{(t)} x^{(t)^T} A$$

It can be shown that $R^{(k)}$ has the least Frobenius norm and least 2-norm of all matrices of the form $A - D$, where $rank(D) \leq k$ (by Eckart-Young Theorem). In other words, the optimal clustering makes the matrix $R^{(k)}$ as small as possible.

The randomized SVD approximation:
Idea: Compute an approximation to the top left singular vectors of A by sampling its columns (with respect to a norm-weighted probability distribution) and constructing a smaller matrix C, then performing truncated SVD on C.

Concretely:

Given $m \times n$ matrix A, denote the probability $p_i = \frac{\left|\left|A^{(i)}\right|\right|^2}{||A||_F^2}$, $i = 1, \ldots, n$. Let $c \leq n$ be the number of columns to sample from A and $k \leq c$ be the number of clusters.

For t=1,…,c:
Pick an integer i from {1,…,n} subject to $\Pr(i) = p_i$
Include $A^{(i)}/\sqrt{cp_i}$ as a column of C

Compute top k left singular vectors of C (denoted $h^{(1)}, \dots, h^{(k)}$)
Stack $h^{(1)}, \dots, h^{(k)}$ as columns of H. Return H and $\lambda_1 = \sigma_1(C), \dots, \lambda_k = \sigma_k(C)$

Proof sketch of major technical results:

Result 1: Error bound of randomized SVD approximation:

If $P = HH^T A$ is a rank k approximation to A constructed using the randomized SVD algorithm, then with probability $1 - \delta$, we have:

$$\|A - P\|_F^2 \leq \|A - A_k\|_F^2 + 2\left(1 + \sqrt{8 \ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$$

$$\|A - P\|_2^2 \leq \|A - A_k\|_2^2 + 2\left(1 + \sqrt{8 \ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$$

Proof sketch for Frobenius norm (2-norm proof omitted):

Letting $h^{(t)}, t = 1, \dots, k$ denote the top k left singular vectors of C and $\sigma_t(C)$ the corresponding singular values, and H be the matrix whose columns are $h^{(t)}$. It can be shown that

$$\|A - HH^T A\|^2{}_F = \|A\|_F^2 - \sum_{t=1}^{k} \left\|A^T h^{(t)}\right\|^2$$

Also, ${h^{(t)}}^T (AA^T - CC^T) h^{(t)}$ is the (t,t) entry of $AA^T - CC^T$. Hence:

$$\sum_{t=1}^{k} \left({h^{(t)}}^T (AA^T - CC^T) h^{(t)}\right)^2 \leq \|AA^T - CC^T\|_F^2$$

Using the fact that $C^T h^{(t)} = \sigma_t(C) h^{(t)}$ and applying Cauchy-Schwartz inequality:

$$\sum_{t=1}^{k} \left(\left\|A^T h^{(t)}\right\|^2 - \sigma_t^2(C)\right) \geq -\sqrt{k}\|AA^T - CC^T\|_F$$

We can use Hoffman-Wielandt inequality and Cauchy Schwartz to get:

$$\sum_{t=1}^{k} \left(\sigma_t^2(C) - \sigma_t^2(A)\right) \geq -\sqrt{k}\|AA^T - CC^T\|_F$$

Summing the two inequalities:

$$\sum_{t=1}^{k} \left(\left\|A^T h^{(t)}\right\|^2 - \sigma_t^2(A)\right) \geq -2\sqrt{k}\|AA^T - CC^T\|_F$$

Invoke Lemma 5, which says that if C is created using the randomized SVD algorithm, then with probability at least $1-\delta$ we have $\|AA^T - CC^T\|_F \leq \frac{1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}}{\sqrt{c}}\|A\|_F^2$. So with high probability, we have:

$$\sum_{t=1}^{k}\left\|A^T h^{(t)}\right\|^2 \geq \sum_{t=1}^{k}\sigma_t^2(A) - 2\left(1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$$

Substitute this inequality into $\|A - HH^T A\|^2{}_F = \|A\|_F^2 - \sum_{t=1}^{k}\left\|A^T h^{(t)}\right\|^2$ and use the fact that $\|A - A_k\|_F^2 = \|A\|_F^2 - \sum_{t=1}^{k}\sigma_t^2(A)$:

$$\|A - HH^T A\|^2{}_F \leq \|A\|_F^2 - \sum_{t=1}^{k}\sigma_t^2(A) + 2\left(1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$$

$$\|A - P\|_F^2 \leq \|A - A_k\|_F^2 + 2\left(1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$$

As desired.

Result 2: Lemma 5

If C is created using the randomized SVD algorithm, then with probability at least $1-\delta$ we have $\|AA^T - CC^T\|_F \leq \frac{1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}}{\sqrt{c}}\|A\|_F^2$.

Proof sketch:

Let $F(i_1, \ldots, i_c) = \|AA^T - CC^T\|_F^2$, where $i_1, \ldots, i_c$ are the indices of columns of A sampled independently.

Consider:

$$CC^T = \sum_{t=1}^{c}\frac{A^{(i_t)}(A^T)_{(i_t)}}{cp_{i_t}}$$

For t=1,…,c, we consider the (i,j) entry of the summand: $w_t = \left(\frac{A^{(i_t)}(A^T)_{(i_t)}}{cp_{i_t}}\right)_{ij}$. We can compute the expectation of w_t and bound its variance.

$$E(w_t) = \frac{1}{c}(AA^T)_{ij}$$

$$Var(w_t) \leq \sum_{k=1}^{n} \frac{A_{ik}^2 (A^T)_{kj}^2}{c^2 p_k}$$

Which allows us to bound the variance of CC^T: $Var(CC^T)_{ij} \leq c \sum_{k=1}^{n} \frac{A_{ik}^2 \left(A^T\right)_{kj}^2}{c^2 p_k}$.

It can then be shown that $E[\|AA^T - CC^T\|_F^2] \leq \frac{1}{c}\|A\|_F^2$ (write out the expectation as a sum over expectation of squared ij entries, which reduces to a sum of variances of $(CC^T)_{ij}$, which is bounded (shown previously)).

We also know that since $E(|X|) \leq \sqrt{E(X^2)}$ (since Var(X)>=0), we have:

$$E[\|AA^T - CC^T\|_F^2] \leq \frac{1}{\sqrt{c}}\|A\|_F^2$$

We consider a perturbation to the sampled columns and bound the effect this has on F: Change just one of the i_t to $i_t{}'$ and let the new matrix be C'. It can be easily shown: $\|CC^T - C'C'^T\|_F \leq \frac{2}{c}\|A\|_F^2$

Now use triangle inequality:

$$\|CC^T - AA^T\|_F \leq \|C'C'^T - AA^T\|_F + \|CC^T - C'C'^T\|_F \leq \|C'C'^T - AA^T\|_F + \frac{2}{c}\|A\|_F^2$$

Invoke Azuma's inequality (similar to a Chernoff bound) to obtain:

$$\Pr\left(\left|\|CC^T - AA^T\|_F - E[\|CC^T - AA^T\|_F]\right| \leq \lambda\sqrt{c}\frac{2}{\beta c}\|A\|_F^2\right) \geq 1 - 2e^{-\lambda^2/2}$$

Using $\delta = 2e^{-\frac{\lambda^2}{2}}$ (or $\lambda = \sqrt{2\ln\left(\frac{2}{\delta}\right)}$, we see that for all $\delta > 0$, with probability at least $1-\delta$, we have $\|CC^T - AA^T\|_F \leq \lambda\sqrt{c}\frac{2}{\beta c}\|A\|_F^2 + E[\|CC^T - AA^T\|_F] \leq \lambda\sqrt{c}\frac{2}{\beta c}\|A\|_F^2 + \frac{1}{\sqrt{c}}\|A\|_F^2 = \frac{1+\sqrt{8\ln\left(\frac{2}{\delta}\right)}}{\sqrt{c}}\|A\|_F^2$

Result 3: Lemma 6

The p_i's specified in the randomized SVD algorithm minimize $E[\|CC^T - AA^T\|_F^2] = \frac{1}{c}\sum_{k=1}^{n}\frac{1}{p_k}\left\|A^{(k)}\right\|^4 - \frac{1}{c}\|AA^T\|_F^2$. Since the second term is constant, we only need to minimize the first term, so consider the function $f(p_1, \ldots, p_n) = \frac{1}{c}\sum_{k=1}^{n}\frac{1}{p_k}\left\|A^{(k)}\right\|^4$. Minimizing f subject to $\sum p_k = 1$ gives the desired result.

Result 4: Efficiency of randomized SVD algorithm

For a $m \times n$ matrix A, suppose we use the algorithm to return P. Ensuring with high probability that $\|A - P\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2$ and $\|A - P\|_2^2 \leq \|A - A_k\|_2^2 + \epsilon\|A\|_2^2$ requires a runtime complexity of $O\left(\frac{k^2}{\epsilon^4}m + \frac{k^3}{\epsilon^6}\right)$.

If we only want to satisfy $\|A - P\|_2^2 \leq \|A - A_k\|_2^2 + \epsilon\|A\|_2^2$ (with high probability), then we need $O\left(\frac{1}{\epsilon^4}m + \frac{1}{\epsilon^6}\right)$.

Proof sketch:
Computing $C^T C$, its SVD, and H requires $O(c^2m + c^3 + cmk) = O(c^2m + c^3)$ since $k \leq c$. We note that the requirement is actually $\|A - P\|_F^2 \leq \|A - A_k\|_F^2 + 2\left(1 + \sqrt{8\ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}\|A\|_F^2$, so, taking $\epsilon = 2\left(1 + \sqrt{8\ln\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{c}}$, we see that $c \sim \frac{k}{\epsilon^2}$. So the total runtime complexity is $O(c^2m + c^3) = O\left(\frac{k^2}{\epsilon^4}m + \frac{k^3}{\epsilon^6}\right)$.

A similar argument can be applied in the case of $\|A - P\|_2^2 \leq \|A - A_k\|_2^2 + \epsilon\|A\|_2^2$ to obtain $O\left(\frac{1}{\epsilon^4}m + \frac{1}{\epsilon^6}\right)$.

Result 5: Accuracy of CCP approximation:
The proposed algorithm returns a 2-approximation for DCP.

Proof sketch:
Let Z_A^{DCP} be the optimal value of the DCP problem with dataset A. The continuous relaxation of the problem leads to lower error, so:

$$Z_A^{DCP} \geq \sum_{i=1}^{m} \left\|A_{(i)} - \bar{A}_{(i)}\right\|^2$$

Let $\hat{B} = \{\hat{B}^{(1)}, \dots, \hat{B}^{(k)}\}$ be the optimal solution to the DCP and $\bar{B}$ be the projection of points in $\hat{B}$ to V. Then:

$$Z_A^{DCP} = \sum_{i=1}^{m} dist^2\left(A_{(i)}, \hat{B}\right) \geq \sum_{i=1}^{m} dist^2\left(\bar{A}_{(i)}, \bar{B}\right)$$

$$2Z_A^{DCP} \geq \sum_{i=1}^{m}\left(\left\|A_{(i)} - \bar{A}_{(i)}\right\|^2 + dist^2\left(\bar{A}_{(i)}, \bar{B}\right)\right) = \sum_{i=1}^{m} dist^2\left(A_{(i)}, \bar{B}\right) = f_A(\bar{B})$$

Problem 2: Hardness of k-center

We note that the VC* optimization problem can be converted to a decision problem: "Is there a vertex cover V' such that $|V'| = k$?" To optimize $|V'|$, we iteratively solve this decision problem for decreasing values of k until it is impossible to go smaller.

Note also that the k-center optimization problem can also be reframed to a decision problem: "Does the k-center problem admit a cost of C?"

We want to reduce this decision problem to a k-center problem. Consider an instance of VC* $\phi = (G = (V, E), k)$. Now construct a new set of vertices $W = \{v_e | e \in E\}$, i.e. create a new vertex for every edge in G. Now consider the metric on $V \cup W$ for the corresponding k-center problem:

$$\rho(u, v) = 1 \text{ if } (u, v) \in E$$
$$\rho(u, v_e) = \rho(v, v_e) = 1 \text{ if } v_e \in W \text{ and } e = (u, v) \in E$$
$$\rho(u, v) = 2 \text{ otherwise}$$

We note that this is a valid metric since the triangle inequality is satisfied.

We now prove that if we can find a VC $|V'| = k$, then we can achieve the minimum cost of 1 in the corresponding $D(\phi)$ k-center problem (taking V' as the cluster centers).

Consider any $v \in V \cup W$. If $v \in V'$, then trivially $\rho(v, V') = 0$. Now consider if $v \in (V - V') \cup W$. There are two possible sub-cases to consider. First, if $v \in V - V'$, since V' is a VC of V, we must have $\rho(v, V') = 1$ (since a VC must cover all the edges of G). Note that, WLOG, we do not consider isolated vertices (since they can be included as cluster centers). If $v \in W$, we can denote $v = v_e$, where $e = (u, w) \in E$ for some vertices $u, w \in V$. Since $\rho(u, v_e) = \rho(w, v_e) = 1$ and at least one of u, w is in V', we have $\rho(v, V') = 1$. Hence, since every vertex not in the cover is at most distance 1 away from the cluster center, $D(\phi)$ admits the minimum cost of 1. Now we prove the other direction: If $D(\phi)$ admits the minimum cost of 1, then $\phi = (G, k)$ returns true.

Let S be the set of k cluster centers that induce the cost of 1 over $V \cup W$. While S contains elements in W, iteratively remove from S the element in W and replace it with an element in $V - V'$ (if such an element exists). The resultant S has $|S| \leq k$, with the inequality holding if we ran out of elements in $V - V'$ to replace with. Note that this S is also a VC of V since $\rho(v, S) \leq 1 \quad \forall v \in V$. Hence $\phi = (G, k)$ returns true.

Hence, $\phi = (G = (V, E), k)$ returns true if and only if $D(\phi)$ admits a cost of 1. So the Vertex-Cover* problem reduces to the k-center problem:

$$\text{VC}^* \leq_p k\_center$$

Since VC* is known to be NP-hard, k-center is NP-hard.

Problem 3: Coverings and packings

3i.

(X, ρ) is compact, so by definition of compactness, for every open cover $\bigcup_\alpha G_\alpha \supset X$, there is a finite subcover $\bigcup_{k=1}^{N} G_k$. In particular, take $G_\alpha = B_\epsilon(x_\alpha)$ (the open ball of radius ϵ around x_α, where the x_α's are all the points in X. Clearly, $\bigcup_\alpha G_\alpha$ is an open cover. By compactness, only N (TBA) of these open balls are enough to cover X.

We can construct Y as follows:

Let $Y = \emptyset$
While X is not yet covered by $\bigcup_{x \in Y} B_\epsilon(x)$
    Pick an $x \in X$ that is not yet in the cover. i.e. $\rho(x, x_i) > \epsilon, \ \forall x_i \in Y$
    $Y \leftarrow Y \cup \{x\}$

Since X is compact, this loop should terminate with a finite subcover, giving us $Y = \{x_1, \dots, x_N\}$ for some N. For this Y, we see that, by construction:

$$\rho(x_i, x_j) > \epsilon \ \forall i \neq j$$

and $\forall x \in X, \exists x_i \in Y$ such that $\rho(x, x_i) < \epsilon$ (by definition of subcover). With the same Y, we can now consider closed balls instead so that $\forall x \in X, \exists x_i \in Y$ s. t. $\rho(x, x_i) \leq \epsilon$.

Hence, this Y is both an epsilon-cover and an epsilon-packing.

3ii.

Proof of right inequality:

As was also seen in the previous part, a maximal $\frac{\epsilon}{2}$-packing p is also an $\frac{\epsilon}{2}$-cover. This is because if p is not an $\frac{\epsilon}{2}$-cover, I can simply take $x \in X$ not in the cover and append x to p and p would still be an $\frac{\epsilon}{2}$-packing. The fact I can do this contradicts the fact that p is a **maximal** $\frac{\epsilon}{2}$-packing, so p must also be an $\frac{\epsilon}{2}$-cover. Then we must have $|p| = P_{\frac{\epsilon}{2}}(X) \geq N_{\frac{\epsilon}{2}}(X)$, since $N_{\frac{\epsilon}{2}}(X)$ is the size of the **minimal** $\frac{\epsilon}{2}$-cover. This proves the right inequality.

Proof of left inequality:

Suppose there is a maximal ϵ-packing and minimal $\frac{\epsilon}{2}$-cover such that $P_\epsilon(X) > N_{\frac{\epsilon}{2}}(X)$. By pigeonhole principle, this means that there are at least two elements in the packing (call them p_1, p_2) that belong to the same $\frac{\epsilon}{2}$-ball. But this contradicts the fact that $\rho(p_1, p_2) > \epsilon$. Hence, we must have that $P_\epsilon(X) \leq N_{\frac{\epsilon}{2}}(X)$, as desired.

3iii.

Let $X = B^d(0,1)$. Let's try to calculate $P_\epsilon(X)$. Since an ϵ-packing consists of elements that are at least ϵ distance apart, this means that each element of the packing must have non-overlapping balls of radius $\frac{\epsilon}{2}$. So estimating $P_\epsilon(X)$ amounts to trying to fill up the volume of $B^d(0,1)$ with volume of smaller balls $B^d\left(0,\frac{\epsilon}{2}\right)$. Hence:

$$P_\epsilon(X) \cong \frac{vol\left(B^d(0,1)\right)}{vol\left(B^d\left(0,\frac{\epsilon}{2}\right)\right)} = \frac{vol\left(B^d(0,1)\right)}{\left(\frac{\epsilon}{2}\right)^d vol\left(B^d(0,1)\right)} = \left(\frac{\epsilon}{2}\right)^{-d} = \left(\frac{2}{\epsilon}\right)^d$$

Similarly:

$$P_{\frac{\epsilon}{2}}(X) = \left(\frac{\epsilon}{4}\right)^{-d} = \left(\frac{4}{\epsilon}\right)^d$$

So we have:

$$\left(\frac{2}{\epsilon}\right)^d \leq N_{\frac{\epsilon}{2}}(X) \leq \left(\frac{4}{\epsilon}\right)^d$$

Or:

$$\left(\frac{1}{\epsilon}\right)^d \leq N_\epsilon(X) \leq \left(\frac{2}{\epsilon}\right)^d$$

3iv.

Let x^* be the x that achieves $\sigma_{max}(A)$, i.e. $x^* = \text{argmax}_{x\in R^n, ||x||=1}||Ax||$. Then, let $x \in C$ be the nearest point to x^* that's in the cover C. Then:

$$\sigma_{max}(A) - \sigma_C(A) = \left|||Ax^*|| - ||Ax||\right| = \left|\left|||Ax^*|| - ||Ax||\right|\right| \leq \left||A(x^* - x)|\right| \leq \left||A|\right||x^* - x|$$
$$\leq \left||A|\right|\epsilon$$

Where we used the inverse triangle inequality and applied the fact that $\left||x^* - x|\right| \leq \epsilon$ (since C is a cover). Note that $\left||A|\right| = \sup_{x\in R^n} \frac{|Ax|}{|x|} = \sup_{x\in R^n, |x|=1} |Ax| = \sigma_{max}(A)$. Hence we have:

$$\sigma_{max}(A) - \sigma_C(A) \leq \sigma_{max}(A)\epsilon$$
$$(1-\epsilon)\sigma_{max}(A) \leq \sigma_C(A) \leq \sigma_{max}(A)$$

An upper bound on $\sigma_{max}(A)$ is thus:

$$\sigma_{max}(A) \leq \frac{\sigma_C(A)}{1-\epsilon}$$

Problem 4: Low-dimensional embeddings from dissimilarity data

4i.
Consider:

$$\langle \alpha_i - \bar{\alpha}, \alpha_j - \bar{\alpha} \rangle = \langle \alpha_i, \alpha_j \rangle - \frac{1}{n}\sum_l \langle \alpha_l, \alpha_j \rangle - \frac{1}{n}\sum_m \langle \alpha_i, \alpha_m \rangle + \frac{1}{n^2}\sum_{l,m} \langle \alpha_l, \alpha_m \rangle$$

Note that: $H_{mp} = \delta_{mp} - \frac{1}{n}$, where δ_{mp} is the Kronecker delta.

Then consider:

$$(DH)_{lp} = \sum_m D_{lm}H_{mp} = \sum_m ||\alpha_l - \alpha_m||^2 \left(\delta_{mp} - \frac{1}{n}\right) = \left|\left|\alpha_l - \alpha_p\right|\right|^2 - \frac{1}{n}\sum_m ||\alpha_l - \alpha_m||^2$$

Note that $H^T = H$ since H is symmetric. Then:

$$-\frac{1}{2}(H^T DH)_{ij} = -\frac{1}{2}\sum_l (H^T)_{il}(DH)_{lj} = -\frac{1}{2}\left[(DH)_{ij} - \frac{1}{n}\sum_l (DH)_{lj}\right]$$

$$= -\frac{1}{2}\left[\left|\left|\alpha_i - \alpha_j\right|\right|^2 - \frac{1}{n}\sum_m ||\alpha_i - \alpha_m||^2 - \frac{1}{n}\sum_l \left|\left|\alpha_l - \alpha_j\right|\right|^2 + \frac{1}{n^2}\sum_{lm} ||\alpha_l - \alpha_m||^2\right]$$

$$= -\frac{1}{2}\left[\langle \alpha_i, \alpha_i \rangle + \langle \alpha_j, \alpha_j \rangle - 2\langle \alpha_i, \alpha_j \rangle - \frac{1}{n}\sum_l \left(\langle \alpha_l, \alpha_l \rangle + \langle \alpha_j, \alpha_j \rangle - 2\langle \alpha_l, \alpha_j \rangle\right)\right.$$
$$\left. - \frac{1}{n}\sum_m \left(\langle \alpha_m, \alpha_m \rangle + \langle \alpha_i, \alpha_i \rangle - 2\langle \alpha_m, \alpha_i \rangle\right.\right.$$
$$\left.\left. - \frac{1}{n}\sum_m (\langle \alpha_m, \alpha_m \rangle + \langle \alpha_l, \alpha_l \rangle - 2\langle \alpha_m, \alpha_l \rangle)\right)\right]$$

After some cross cancellations we get:

$$= -\frac{1}{2}\left[-2\langle \alpha_i, \alpha_j \rangle + \frac{2}{n}\sum_l \langle \alpha_l, \alpha_j \rangle + \frac{2}{n}\sum_m \langle \alpha_m, \alpha_i \rangle - \frac{2}{n^2}\sum_{lm} \langle \alpha_l, \alpha_m \rangle\right]$$
$$= \langle \alpha_i, \alpha_j \rangle - \frac{1}{n}\sum_l \langle \alpha_l, \alpha_j \rangle - \frac{1}{n}\sum_m \langle \alpha_i, \alpha_m \rangle + \frac{1}{n^2}\sum_{l,m} \langle \alpha_l, \alpha_m \rangle$$

Hence, comparing this with $\langle \alpha_i - \bar{\alpha}, \alpha_j - \bar{\alpha} \rangle$, we see that:

$$-\frac{1}{2}(H^T DH)_{ij} = \langle \alpha_i - \bar{\alpha}, \alpha_j - \bar{\alpha} \rangle$$

As desired.

4ii:

$$B_{ij} = \langle \beta_i, \beta_j \rangle$$

So for any non-zero vector v:

$$v^T B v = \sum_{l,m} v_l \langle \beta_l, \beta_m \rangle v_m = \langle \sum_l v_l \beta_l \, , \sum_m v_m \beta_m \rangle \geq 0$$

Where $\beta_i = \alpha_i - \bar{\alpha}$. Hence, the Gram matrix B is positive semidefinite, which means it is symmetric and diagonalizable with non-negative eigenvalues:

$$B = Q \Lambda Q^T = Q \Lambda^{1/2} \Lambda^{1/2} Q^T = (X - E[X])^T (X - E[X])$$

So we can take the rows of $Q\Lambda^{1/2} \in \mathbb{R}^{n \times n}$ as the embedding. Let's prove that the rows are a perfect (isometric) embedding:

Consider the quantity:

$$\begin{aligned}
&\left|\left| \left(Q\Lambda^{\frac{1}{2}}\right)_i - \left(Q\Lambda^{\frac{1}{2}}\right)_j \right|\right|^2 \\
&= (Q\Lambda Q^T)_{ii} + (Q\Lambda Q^T)_{jj} - 2(Q\Lambda Q^T)_{ij} \\
&= B_{ii} + B_{jj} - 2B_{ij} \\
&= \langle \alpha_i - \bar{\alpha}, \alpha_i - \bar{\alpha} \rangle + \langle \alpha_j - \bar{\alpha}, \alpha_j - \bar{\alpha} \rangle - 2\langle \alpha_i - \bar{\alpha}, \alpha_j - \bar{\alpha} \rangle \\
&= \langle \alpha_i, \alpha_i \rangle - 2\langle \alpha_i, \bar{\alpha} \rangle + \langle \bar{\alpha}, \bar{\alpha} \rangle + \langle \alpha_j, \alpha_j \rangle - 2\langle \alpha_j, \bar{\alpha} \rangle + \langle \bar{\alpha}, \bar{\alpha} \rangle \\
&\qquad\qquad - 2\big(\langle \alpha_i, \alpha_j \rangle - \langle \alpha_j, \bar{\alpha} \rangle - \langle \alpha_i, \bar{\alpha} \rangle + \langle \bar{\alpha}, \bar{\alpha} \rangle\big) \\
&= \langle \alpha_i, \alpha_i \rangle - 2\langle \alpha_i, \alpha_j \rangle + \langle \alpha_j, \alpha_j \rangle \\
&= \left|\left| \alpha_i - \alpha_j \right|\right|^2 = D_{ij}
\end{aligned}$$

So the rows of $Q\Lambda^{1/2}$ are indeed a perfect isometric embedding, as desired.

4iii. (Ignored, as per TA's instruction)

Problem 5: Effects of a Gaussian Random Matrix

Consider:

$$(Gv)_i = \sum_{j=1}^{D} G_{ij} v_j$$

Note that $G_{ij} \sim N(0, \frac{1}{d})$ iid, so $G_{ij} v_j \sim N\left(0, \frac{v_j^2}{d}\right)$. Due to independence, the variance of the sum is the sum of the variance, so:

$$(Gv)_i = \sum_{j=1}^{D} G_{ij} v_j \sim N\left(0, \frac{\sum_{j=1}^{D} v_j^2}{d}\right) = N\left(0, \frac{1}{d}\right)$$
$$\forall i = 1, \dots d$$

So, let $\left|\left|Gv\right|\right|^2 = \sum_{i=1}^{d} (Gv)_i^2 = \frac{1}{d} Y$, where $Y \sim \chi_d^2$.

From the result in HW0, we see that:

$$P\left[\left|\left|Gv\right|\right|^2 > (1+\epsilon)^2\right] = P\left[\frac{1}{d} Y > (1+\epsilon)^2\right] = P[Y > d(1+\epsilon)^2] < e^{-cd\epsilon^2}$$

As desired.