1. Given a graph $G(V, E)$ with each edge (i, j) assigned two weights: the *similarity* between two vertices i, j denoted by s_{ij} and the *difference* between the vertices denoted by d_{ij}. We want to cluster the vertex set of G into disjoint vertex clusters - say $C = (C_1, C_2, .., C_n)$, such that every vertex is in exactly one C_n. Note that some C_k could be empty. Two vertices i, j are said to be *together* in a clustering C if they belong to the same C_k. Otherwise we say i, j are *apart*. The *correlation* of a clustering C, denoted by $\text{Cor}(C)$ is the sum of the $s_{i,j}$ values of all i, j that are **together**. The *anti-correlation* of C, $\text{Acor}(C)$ is the sum of the differences d_{ij} of all i, j which are **apart**. We want to find a clustering C that **maximizes** the sum $\text{Cor}(C)+\text{Acor}(C)$.

    1. Considering the two trivial clusterings: 1). C that clusters all n vertices into a single cluster and 2). C' which puts each vertex into a different cluster. Show that one of C or C' must be within a factor of $\frac{1}{2}$ from the optimal. Thus the problem is somewhat trivially 0.5-approximable. In the following, we develop an SDP formulation.
    2. Given a clustering C of vertices, let v_i be a unit vector such that $v_i(k) = 1$ if and only if vertex i is in cluster k. (Thus each $v_i \in \{e_1, e_2, ..., e_n\}$ where $(e_1, e_2, ..e_n)$ is the standard basis of $\mathbf{R}^n$). Show that the following formulation Maximize $\sum_{i,j} s_{ij}(v_i, v_j) + \sum_{i,j} d_{ij}(1 - (v_i, v_j))$ subject to constraints $v_i \in \{e_1, e_2, ..., e_n\}$ for each i is an exact formulation for the problem.
    3. However, the constraint $v_i \in \{e_1, e_2, ..., e_n\}$ makes the above formulation NP hard. Hence consider the relaxation dropping this constraint and adding constraints $(v_i, v_i) = 1$ for each i. (Thus, we allow v_i to be any unit vector, no longer requiring it to be a standard basis vector). We also add additional constraints $(v_i, v_j) \geq 0$ for all i, j. Show that this is a vector program relaxation of the original problem in that any solution to the original problem will satisfy the new system.

2. In this question, we develop a random hyperplane approximation algorithm for rounding the optimal solution to the vector program developed in the previous question. Let $v_1, v_2, .., v_n$ be the optimal SDP solution values obtained by solving the relaxation. To find an approximate solution, consider **two** random unit vectors r_1 and r_2 in $\mathbf{R}^n$ sampled by normal distribution with mean 0 and variance 1 as discussed in the class. We divide the vertices of G into *four* clusters, C_1, C_2, C_3 and C_4 as the following:

    - Vertex i is added to cluster C_1 if $(v_i, r_1) \geq 0$ and $(v_i, r_2) \geq 0$.
    - Vertex i is added to cluster C_2 if $(v_i, r_1) \geq 0$ and $(v_i, r_2) < 0$.
    - Vertex i is added to cluster C_3 if $(v_i, r_1) < 0$ and $(v_i, r_2) \geq 0$.
    - Vertex i is added to cluster C_4 if $(v_i, r_1) < 0$ and $(v_i, r_2) < 0$.

    The following questions prove that this clustering yields a 0.75 factor approximation for the clustering problem discussed in the previous question. Let $\theta_{ij} = \cos^{-1}(v_i, v_j)$ be the angle between v_i and v_j. Let X_{ij} be a random variable that takes value 1 if vertices i and j are in the *same* cluster, 0 if i and j are in different clusters. You will need the *trigonometric inequalities* $(1 - \frac{\theta}{\pi})^2 \geq \frac{3}{4}\cos\theta$ and $1 - (1 - \frac{\theta}{\pi})^2 \geq \frac{3}{4}(1 - \cos\theta)$ if $0 \leq \theta \leq \frac{\pi}{2}$.

    1. Show that the optimal relaxed solution has value $\sum_{i,j} s_{ij}\cos\theta_{ij} + \sum_{i,j} d_{ij}(1 - \cos\theta_{ij})$
    2. Show that the expected size of the rounded solution is $\sum_{i,j} s_{ij}E(X_{ij}) + \sum_{i,j} d_{ij}(1 - E(X_{ij}))$
    3. Show that the probability that a single random hyperplane separates v_i and v_j is $(1 - \frac{\theta_{ij}}{\pi})$. Argue that the probability that vertices i and j land up in the same cluster is $(1 - \frac{\theta_{ij}}{\pi})^2$. Hence conclude that $E(X_{ij}) = (1 - \frac{\theta_{ij}}{\pi})^2$.
    4. Using the trigonometric inequalities noted above, conclude that the expected value of the clustering obtained by the rounding strategy is $\frac{3}{4}$ factor within the optimal.
    5. Where was the constraint $(v_i, v_j) \geq 0$ used in deriving the approximation guarantee?

3. Consider the NP completeness reduction from 3SAT to CLIQUE discussed in the class. (For each clause we added three vertices, one per literal in the clause, and added edges between literals in different clauses if they were mutually consistent). Suppose ϕ is a boolean formula containing k literals. Let G be the graph constructed by the reduction.

    1. Show that G has a clique of size t if and only if there is a truth assignment to ϕ that satisfies t clauses.
    2. Show that the reduction is a gap preserving reduction from 3SAT to CLIQUE.

4. Given a graph $G(V, E)$, consider the square graph G^2 constructed as follows: $V(G^2) = V \times V$. Two vertices (u_1, v_1) and (u_2, v_2) are adjacent in G^2 if *both* the conditions below are true: 1) u_1, u_2 are either adjacent in G or $u_1 = u_2$ (that is, u_1 and u_2 are at distance at most 1 in G). 2) v_1, v_2 are adjacent in G or $v_1 = v_2$ (that is, v_2 and v_1 are at distance at most 1 in G).

    1. If S is a clique of size k in G, then show that $S' = S \times S = \{(u, v) : u, v \in S\}$ is a clique of size k^2 in G^2.

□

2. Conversely, if S is a clique in G^2, Let S_1 and S_2 be the vertices appearing in each coordinate of S. Argue that S_1 and S_2 must be cliques in G. Let $k = \max\{|S_1|, |S_2|\}$. Show that $k \geq \sqrt{|S|}$
3. Argue that if S is a largest clique in G^2 of size k, then G must have a clique of size at least $\sqrt{k}$
4. From the above, conclude that if there is an algorithm that on an input graph G and max clique size k returns a clique of size at least αk, then by running the algorithm of G^2 and using the observation above, we can construct a clique of size at least $\sqrt{\alpha}k$ in G. (Note that $\sqrt{\alpha} > \alpha$ when $0 < \alpha < 1$).
5. Show that if MAXCLIQUE is approximable upto α for some $\alpha > 0$, then MAXCLIQUE is α approximable for every $\alpha > 0$.

5. Given a graph G consider the problem of finding a maximum Independent set in G. Show that if there exists an approximation algorithm for the problem achieving an α factor approximation for some $\alpha > 0$, then the problem is approximable within α for every $\alpha > 0$. (Hint: G has a clique of size k if and only if $\overline{G}$ has an Independent of size k). □

6. Assume that there is a polynomial time algorithm A that on input boolean formula ϕ produces a 3CNF formula $A(\phi)$ such that if $\phi \in$ SAT, then $A(\phi)$ is satisfiable, whereas if ϕ is not satisfiable, then at least ϵ fraction of clauses of $A(\phi)$ is unsatisfiable. Show that then NP$\subseteq$ PCP$(\log n, 1)$. □