Supporting Lemmas and Theorems

We will first prove a lemma that selecting a maximally distinct set of elements according to a distance metric is *NP-Hard* as is finding a polynomial approximation that guarantees a solution that is better than one-half optimal.

Lemma 0.1 *For a set of elements V, a set $R \subset V$, and a distance metric d let $b(R) = min_{v_1,v_2 \in R} d(v_1, v_2)$. Set $b' = max_{R \subset V, |R|=m} b(R)$. Then finding a set R such that $|R| = m$ and $b(R) = b'$ is* NP-Hard. *Furthermore finding a set R such that $|R| = m$ and $b(R) > \frac{b'}{2}$ is also* NP-Hard.

Proof: We first note that given an undirected graph (V, E) the problem of finding an independent set of size m, that is a subset of m vertices such that there is no edge between any two vertices in the subset, is *NP-Hard* [1]. We will define the distance d to be:

$$d(u,v) = \begin{cases} 0 & \text{if}(u = v) \\ 1 & \text{if}(u \neq v) \wedge ((u,v) \in E) \\ 2 & \text{if}(u \neq v) \wedge ((u,v) \notin E) \end{cases}$$

Note that all requirements of a metric are trivially satisfied except the triangle inequality, $d(x,z) \leq d(x,y) + d(y,z)$. To verify the triangle inequality also holds, note that for any two elements, x and z, the possible values of $d(x,z)$ are 0, 1, and 2. If $d(x,z) = 0$ then the triangle inequality is trivially satisfied since d is non-negative. If $d(x,z) = 1$ and the triange inequality was not satisfied, we would have $d(x,y) = 0$ and $d(y,z) = 0$, which would imply $x = z$ and $d(x,z) = 0$ contradicting $d(x,z) = 1$. If $d(x,z) = 2$ and the triangle inequality was not satisfied, then we have $d(x,y) + d(y,z) \leq 1$, which means either $d(x,y) = 0$ or $d(y,z) = 0$. WLOG assume $d(x,y) = 0$, then we have $x = y$ and $d(y,z) = d(x,z) = 2$, which gives a contradiction. Thus the triangle inequality is satisfied in all cases.

We observe that if R is a subset of V of size m such that $b(R) = b'$ and $b(R) > 1$, then the subset of vertices of V which are also elements of R must form an independent set of size m. Furthermore we observe that if $b(R) \leq 1$, then there is no independent set in V of size m. We have thus reduced independent set to being an instance of our problem hence finding an optimal solution to our problem is *NP-Hard*. Furthermore if we could find a subset R of V of size m for which it is guaranteed that $b(R) > \frac{b'}{2}$ in polynomial time, then by the same reduction we could solve independent set in polynomial time, and thus the problem of finding an approximation which guarantees $b(R) > \frac{b'}{2}$ in polynomial time is *NP-Hard*. ∎

In Section 2 we defined $gm(a,b) = 1 - \rho(a,b)$ where ρ is the correlation coefficient, and noted that it does not satisfy the triangle inequality, but does satisfy a generalization of it. We will now prove Lemma 2.1.

Lemma 2.1 $gm(x,z) \leq 2(gm(x,y) + gm(y,z))$

Proof: Given a vector $a = (a_1, ..., a_n)$ we denote the mean normalized vector

$$a' = \left(a_1 - \frac{1}{n}\sum_{i=1}^{n} a_i, ..., a_n - \frac{1}{n}\sum_{i=1}^{n} a_i \right) \tag{1}$$

Let us denote the mean normalized angle between two vectors a and b in radians by $\theta_{a'b'}$ where $0 \leq \theta_{a'b'} \leq \pi$. Note that we are defining the angle between two vectors to be the minimum angle between the vectors, thus the angle will never be greater than π. We note that in general $gm(a,b) = 1 - \rho(a,b) = 1 - \cos(\theta_{a'b'})$

since

$$\rho(a,b) = \frac{\sum_{i=1}^{n}\left(\left(a_i - \frac{1}{n}\sum_{i=1}^{n} a_i\right)\left(b_i - \frac{1}{n}\sum_{i=1}^{n} b_i\right)\right)}{\left(\sum_{i=1}^{n}\left(a_i - \frac{1}{n}\sum_{i=1}^{n} a_i\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}\left(b_i - \frac{1}{n}\sum_{i=1}^{n} b_i\right)^2\right)^{\frac{1}{2}}} \tag{2}$$

$$= \frac{a' \cdot b'}{||a'|| \times ||b'||} \tag{3}$$

$$= cos(\theta_{a',b'}) \tag{4}$$

We first consider the case that $gm(x,y) + gm(y,z) = 0$.

$$gm(x,y) + gm(y,z) = 0 \tag{5}$$

$$\Rightarrow gm(x,y) = 0 \wedge gm(y,z) = 0 \tag{6}$$

$$\Rightarrow \cos(\theta_{x'y'}) = 1 \wedge \cos(\theta_{y'z'}) = 1 \tag{7}$$

$$\Rightarrow \theta_{x'y'} = 0 \wedge \theta_{y'z'} = 0 \tag{8}$$

$$\Rightarrow \theta_{x'z'} = 0 \tag{9}$$

$$\Rightarrow gm(x',z') = 0 \tag{10}$$

and thus the lemma is satisfied for this case.

We will now consider the case that $gm(x',y') + gm(y',z') \neq 0$ and show the function

$$\frac{gm(x,z)}{gm(x,y) + gm(y,z)} = \frac{1 - \cos(\theta_{x'z'})}{2 - \cos(\theta_{x'y'}) - \cos(\theta_{y'z'})} \tag{11}$$

is bounded by 2. We note that cosine is a monotone decreasing function on the interval $[0,\pi]$ and thus for any fixed values of $\theta_{x'y'}$ and $\theta_{y'z'}$ Equation (11) will be greatest when the angle between x' and z' is the largest possible. For any fixed values of $\theta_{x',y'}$ and $\theta_{y',z'}$, $\theta_{x',z'}$ can be no greater than:

$$min(\theta_{x'y'} + \theta_{y'z'}, 2\pi - (\theta_{x'y'} + \theta_{y'z'})) \tag{12}$$

If we fix the angle between x' and z' we observe that since the numerator is always non-negative and the denominator is always positive, that Equation (11) will reach its maximum when the denominator is minimized or equivalently when $\cos(\theta_{x'y'}) + \cos(\theta_{y'z'})$ is maximized.
Thus to bound Equation (11) we need to consider the two cases:
Case 1: The angle between x' and z' is $(\theta_{x'y'} + \theta_{y'z'})$
We thus have the upper bound on Equation (11) of:

$$\frac{1 - \cos(\theta_{x'y'} + \theta_{y'z'})}{2 - \cos(\theta_{x'y'}) - \cos(\theta_{y'z'})} \tag{13}$$

For any fixed value of $C = \theta_{x'y'} + \theta_{y'z'}$ the ratio will thus be maximized when $\cos(\theta_{x'y'}) + \cos(C - \theta_{x'y'})$ is maximized. We consider

$$\frac{\partial}{\partial\theta_{x'y'}}(\cos(\theta_{x'y'}) + \cos(C - \theta_{x'y'})) = -\sin(\theta_{x'y'}) + \sin(C - \theta_{x'y'}) \tag{14}$$

and observe that it is equal to 0 if and only if $\sin(\theta_{x'y'}) = \sin(C - \theta_{x'y'})$ which is true only if either of these equations hold:

$$\theta_{x'y'} = \pi - (C - \theta_{x'y'}) + 2\pi \times k \tag{15}$$

$$\theta_{x'y'} = C - \theta_{x'y'} + 2\pi \times k \tag{16}$$

where k is an integer. The first equation implies $C = \pi(1 + 2k)$ which is true for $C \in [0, \pi]$ only if $k = 0$ and $C = \pi$, in which case Equation (13) is:

$$\frac{1 - \cos(\pi)}{2 - \cos(\theta_{x'y'}) - \cos(\pi - \theta_{x'y'})} = \frac{2}{2} = 1 \tag{17}$$

Assuming $C \neq \pi$ a root otherwise occurs only if $\theta_{x'y'} = \frac{C}{2} + \pi k$ and $\theta_{y'z'} = \frac{C}{2} - \pi k$.
For $\theta_{x'y'}$ and $\theta_{y'z'}$ to both be in $[0, \pi]$ it must be the case that $k = 0$ and thus $\theta_{x'y'} = \theta_{y'z'} = \frac{C}{2}$. We note that we have a maximum when $\theta_{x'y'} = \theta_{y'z'} = \frac{C}{2}$ since

$$\frac{\partial^2}{\partial \theta_{x'y'}^2}(\cos(\theta_{x'y'}) + \cos(C - \theta_{x'y'})) \tag{18}$$

$$= \frac{\partial}{\partial \theta_{x'y'}}(-\sin(\theta_{x'y'}) + \sin(C - \theta_{x'y'})) \tag{19}$$

$$= -\cos(\frac{C}{2}) - \cos(C - \frac{C}{2}) \tag{20}$$

$$= -2\cos(\frac{C}{2}) < 0 \tag{21}$$

The last inequality follows from the fact that $\frac{C}{2}$ must be between $[0, \frac{\pi}{2})$ and thus $\cos(\frac{C}{2}) > 0$.
If $\theta_{x'y'} = \theta_{y'z'}$ the maximum value of Equation (13) is thus the same as the maximum value of

$$\frac{1 - \cos(2\theta_{x'y'})}{2 - 2\cos(\theta_{x'y'})} \tag{22}$$

At $\theta_{x'y'} = \pi$ the value of Equation (22) is 0. At the other end of the $[0, \pi]$ interval we have a maximum of 2 by applying l'Hopital's rule twice:

$$\lim_{\theta_{x'y'} \to 0} \frac{1 - \cos(2\theta_{x'y'})}{2 - 2\cos(\theta_{x'y'})} = \lim_{\theta_{x'y'} \to 0} \frac{\sin(2\theta_{x'y'})}{\sin(\theta_{x'y'})} = \lim_{\theta_{x'y'} \to 0} \frac{2\cos(2\theta_{x'y'})}{\cos(\theta_{x'y'})} = 2 \tag{23}$$

Equation (22) does not reach a local maximum on $(0, \pi)$ since

$$\frac{\partial}{\partial \theta_{x'y'}} \frac{1 - \cos(2\theta_{x'y'})}{2 - 2\cos(\theta_{x'y'})} \tag{24}$$

$$= \frac{2\sin(2\theta_{x'y'})(2 - 2\cos(\theta_{x'y'})) - 2(1 - \cos(2\theta_{x'y'}))\sin(\theta_{x'y'})}{(2 - 2\cos(\theta_{x'y'}))^2} \tag{25}$$

$$= \frac{2\sin(\theta_{x'y'})\cos(\theta_{x'y'})(2 - 2\cos(\theta_{x'y'})) - (1 - \cos(2\theta_{x'y'}))\sin(\theta_{x'y'})}{2(1 - \cos(\theta_{x'y'}))^2} \tag{26}$$

which equals 0, if and only if the numerator equals 0. The numerator is equivalent to

$$= \sin(\theta_{x'y'})[2\cos(\theta_{x'y'})(2 - 2\cos(\theta_{x'y'})) - (1 - \cos(2\theta_{x'y'}))] \tag{27}$$

$$= \sin(\theta_{x'y'})[4\cos(\theta_{x'y'}) - 4\cos(\theta_{x'y'})^2 - 1 + \cos(2\theta_{x'y'})] \tag{28}$$

$$= \sin(\theta_{x'y'})[4\cos(\theta_{x'y'}) - 4\cos(\theta_{x'y'})^2 - 1 + \cos(\theta_{x'y'})^2 - \sin(\theta_{x'y'})^2] \tag{29}$$

$$= -2\sin(\theta_{x'y'})[\cos(\theta_{x'y'})^2 - 2\cos(\theta_{x'y'}) + 1] \tag{30}$$

$$= -2\sin(\theta_{x'y'})(\cos(\theta_{x'y'}) - 1)^2 \tag{31}$$

and thus does not have any roots on $(0, \pi)$
We finally note that if $C = 0$ Equation (11) is 0.

Case 2 *The angle between x' and z' is* $(2\pi - \theta_{x'y'} - \theta_{y'z'})$
We thus have the upper bound on Equation (11) of:

$$\frac{1 - \cos(2\pi - \theta_{x'y'} - \theta_{y'z'})}{2 - \cos(\theta_{x'y'}) - \cos(\theta_{y'z'})} \tag{32}$$

Fix $C = (2\pi - \theta_{x'y'} - \theta_{y'z'})$ and note that $\theta_{y'z'} = 2\pi - C - \theta_{x'y'}$.
The ratio will thus be maximized when $\cos(\theta_{x'y'}) + \cos(2\pi - C - \theta_{x'y'})$ is maximized. We consider

$$\frac{\partial}{\partial \theta_{x'y'}}(\cos(\theta_{x'y'}) + \cos(2\pi - C - \theta_{x'y'})) \tag{33}$$

$$= \frac{\partial}{\partial \theta_{x'y'}}(\cos(\theta_{x'y'}) + \cos(C + \theta_{x'y'})) \tag{34}$$

$$= -\sin(\theta_{x'y'}) - \sin(C + \theta_{x'y'}) \tag{35}$$

and observe that it is equal to 0 if and only if $\sin(\theta_{x'y'}) = \sin(-C - \theta_{x'y'})$ which is true only if either of these equations hold

$$\theta_{x'y'} = \pi - (-C - \theta_{x'y'}) + 2\pi \times k \tag{36}$$

$$\theta_{x'y'} = -C - \theta_{x'y'} + 2\pi \times k \tag{37}$$

where k is an integer. The first of the two equations is satisfied only if $C = (-2k - 1)\pi$ which occurs for $C \in [0, \pi]$ only if $C = \pi$. If $C = \pi$, our upper bound on Equation 11 is

$$\frac{1 - \cos(\pi)}{2 - \cos(\theta_{x'y'}) - \cos(\pi - \theta_{x'y'})} = \frac{2}{2} = 1 \tag{38}$$

Suppose $C \neq \pi$, we also have a root if $\theta_{x'y'} = \pi \times k - \frac{C}{2}$, to have $\theta_{x'y'} \in [0, \pi]$ it must be the case that $\theta_{x'y'} = \pi - \frac{C}{2}$ and then $\theta_{y'z'} = 2\pi - \theta_{x'y'} - C = 2\pi - (\pi - \frac{C}{2}) - C = \pi - \frac{C}{2}$. We note that this root is a local minimum since

$$\frac{\partial^2}{\partial \theta_{x'y'}^2}(\cos(\theta_{x'y'}) + \cos(2\pi - C - \theta_{x'y'})) \tag{39}$$

$$= \frac{\partial}{\partial \theta_{x'y'}}(-\sin(\theta_{x'y'}) - \sin(C + \theta_{x'y'})) \tag{40}$$

$$= -\cos(\pi - \frac{C}{2}) - \cos(C + (\pi - \frac{C}{2})) \tag{41}$$

$$= -\cos(\pi - \frac{C}{2}) - \cos(\pi + \frac{C}{2}) \tag{42}$$

$$= 2\cos(\frac{C}{2}) > 0 \tag{43}$$

The last inequality follows from C being between $[0, \pi)$ and thus $\cos(\frac{C}{2}) > 0$.
We also observe in this case also that if $C = 0$ Equation (11) is 0.
Thus in this case Equation (11) is bounded by 1, and overall Equation (11) is bounded by 2. ∎

We further remark that the proof actually gave us a stronger bound on $gm(x,z)$, given the values of $gm(x,y)$ and $gm(y,z)$, we have that

$$g(x,z) \leq 1 - \cos(\arccos(1 - gm(x,y)) + \arccos(1 - gm(y,z))) \leq 2 \times (g(x,y) + g(y,z)) \tag{44}$$

Also in Section 2 we remarked that since gm does not satisfy the triangle inequality Theorem 2.1 is not applicable, however we could still guarantee that the solution of the approximation algorithm presented in Figure 2 was no worse than $\frac{1}{4}$ optimal. We will prove next the result that if a distance function d satisfies all the properties of a metric, except only satisfies a generalized form of the triangle inequality, in which $d(a,c) \leq Y \times (d(a,b) + d(b,c))$ for a fixed $Y \geq 1$ that the approximation algorithm gives a solution that is no worse than $\frac{1}{2Y}$ of optimal. The proof is a simple generalization of Theorem 2.1.

Theorem 0.1 *Let our distance function d satisfy a generalized triangle inequality $d(a,c) \leq Y \times (d(a,b) + d(b,c))$ for a fixed $Y \geq 1$ and otherwise all properties of a distance metric. Let $R' \subset P$ be the set of profiles that maximizes*

$$max_{R \subset P, |R|=m} min_{p_1,p_2 \in R} d(p_1,p_2) \tag{45}$$

Let $R \subset P$ be the set of profiles returned by the approximation algorithm in Figure 2, then $b(R) \geq \frac{b(R')}{2Y}$.

Proof: Set $b' = b(R')$ (b' is the optimal distance) and $b = b(R)$ (b is the distance returned by our algorithm). Let $\{r'_1, r'_2, ..., r'_{m-1}, r'_m\}$ be the profiles in R' and $\{r_1, r_2, ..., r_{m-1}, r_m\}$ be the profiles in R. Note that for any profile $p \in P$ there exists a profile $r_j \in R$ s.t. $d(p, r_j) \leq b$. If p is one of the profiles in R then this is trivially satisfied. If $p \notin R$ then there must be a profile in R with a distance at most b from p otherwise the greedy algorithm would have selected p from R instead of r_m (we know that the minimum distance b was achieved by the last profile r_m). For each profile in R' we can find its closest profile in R. Next, we consider two possible cases, which are also the only possible cases:
Case 1 - Two different profiles, $r'_i, r'_j \in R'$, are closest to the same profile $r_h \in R$:
We note that $d(r'_i, r_h) \leq b$ and $d(r'_j, r_h) \leq b$ as mentioned above. Using the triangle inequality we get $2b \geq d(r'_i, r_h) + d(r'_j, r_h) \geq \frac{d(r'_i, r'_i)}{Y} \geq \frac{b'}{Y}$ and thus $b(R) \geq \frac{b(R')}{2Y}$.
Case 2 - No two elements in R' are closest to the same element in R:
WLOG let r'_m be the element which is closest to r_m (the last profile added by our algorithm). We next observe that there must exists $i \neq m$ such that $d(r'_m, r_i) \leq b$. This is so because if such a profile r_i did not exist then the greedy algorithm would have selected r'_m instead of r_m. Let r'_i be the profile from R' closest to r_i, then $d(r'_i, r_i) \leq b$ since all profiles are within b of a profile selected by the greedy algorithm. We thus have $2b \geq d(r'_m, r_i) + d(r'_i, r_i) \geq \frac{d(r'_i, r'_m)}{Y} \geq \frac{b'}{Y}$ and thus $b(R) \geq \frac{b(R')}{2Y}$.
∎

We further note that assuming $NP \neq P$ no polynomial approximation algorithm can guarantee a result better than $\frac{b'}{2Y}$ in all cases. The proof is a simple modification of the proof of Lemma 0.1:

Lemma 0.2 *For a set of elements V, a set $R \subset V$, and a distance function d that satisfies a generalized triangle inequality, $d(a,c) \leq Y \times (d(a,b) + d(b,c))$ for a fixed $Y \geq 1$ and $\forall a,b,c \in V$, and otherwise all other properties of a distance metric ($\forall a,b \in V\ d(a,b) \geq 0, d(a,a) = 0$, and $d(a,b) = d(b,a)$) let $b(R) = min_{v_1,v_2 \in R} d(v_1,v_2)$. Set $b' = max_{R \subset V, |R|=m} b(R)$. Then finding a set R such that $|R| = m$ and $b(R) = b'$ is* NP-Hard. *Furthermore finding a set R such that $|R| = m$ and $b(R) > \frac{b'}{2Y}$ is also* NP-Hard.

Proof: We first note that given an undirected graph (V, E) the problem of finding an independent set of size m, that is a subset of m vertices such that there is no edge between any two vertices in the subset, is *NP-Hard* [1]. We will define the distance d to be:

$$d(u,v) = \begin{cases} 0 & \text{if}(u = v) \\ 1 & \text{if}(u \neq v) \wedge ((u,v) \in E) \\ 2Y & \text{if}(u \neq v) \wedge ((u,v) \notin E) \end{cases}$$

Note that all requirements of the distance function stated in the lemma are trivially satisfied except the generalized triangle inequality. To verify the generalized triangle inequality also holds note that for any two elements a, c the possible values of $d(a, c)$ are 0, 1, and $2Y$. If $d(a, c) = 0$ then we must have $d(a, b) = d(b, c) = 0$. If $d(a, c) = 1$ or $d(a, c) = 2Y$ then we must have $d(a, b) \geq 1$ and $d(b, c) \geq 1$. In all cases $d(a, c) \leq Y(d(a, b) + d(b, c))$ and thus the generalized triangle inequality is satisfied. We observe that if R is a subset of V of size m such that $b(R) = b'$ and $b(R) > 1$, then the set of vertices V which are also in R must form an independent set of size m. Furthermore we observe that if $b(R) \leq 1$, then there is no independent set in V of size m. We have thus reduced independent set to being an instance of our problem hence finding an optimal solution to our problem is *NP-Hard*. Furthermore if we could find a subset R of V of size m for which it is guaranteed that $b(R) > \frac{b'}{2Y}$ in polynomial time, then by the same reduction we could solve independent set in polynomial time, and thus the problem of finding an approximation which guarantees $b(R) > \frac{b'}{2Y}$ in polynomial time is *NP-Hard*. ∎

References

[1] D.S. (Editor) Hochbaum. *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company, 1997.

Note: This file was updated August 2007. Lemma 0.1 and Lemma 0.2 previously were written stating V was a set of vectors, in this corrected version the Lemma assumes a more general setting. Proof of Lemma 0.1 further clarified August 2008.