

Geometric Embeddings and Graph Partitioning

by Sanjeev Arora, Satish Rao, Umesh Vazirani (ARV)

Instructor: Joseph Cheriyan

Scribe: Tyler Lu

1 Introduction

In these notes, we will focus on approximating the NP-hard problem of finding balanced cuts where we require both sets of cut vertices to be large. Specifically, this is an exposition of the STOC'04 breakthrough result of Arora, Rao and U. Vazirani [1]. The main result is a $O(\sqrt{\log n})$ approximation algorithm for a variety of graph partitioning problems including Sparsest Cut, Edge Expansion and Graph Conductance. Thus improving on the previous best $O(\log n)$ approximation of Leighton and Rao [3].

We will only give a $O(\log^{2/3} n)$ approximation algorithm for balanced cut. However, the tools and techniques developed is the basis of obtaining $O(\sqrt{\log n})$ approximation in a variety of related problems.

2 Problem Model

We begin with the problem setup and a few definitions.

Definition 2.1 (Balanced Cut Problem). Fix constant $0 \leq c' \leq 1$. Given an unweighted graph $G = (V, E)$, let $\delta(S)$ denote the set of edges in the cut of $S \subseteq V$. Find S that

$$\min_{|S|, |V \setminus S| \geq c'n} |\delta(S)|$$

We focus on this problem since the methods developed can be easily adapted to other graph partitioning problems such as Sparsest Cut using well known reductions that preserve approximation ratio up to constant factors.

We will often use the notation $\{e_1, e_2, \dots\}$ to denote a finite set of elements.

2.1 Embedding into ℓ_2^2 Metric

In the ARV approximation algorithm, it works by first mapping the vertices into some high dimensional space $\mathbb{R}^d$ satisfying a special triangle inequality (forming a ℓ_2^2 metric) and requiring the sum of the distances to be large enough (c -spread condition). Why we do this will be clear later.

Definition 2.2 (ℓ_2^2 metric). A set of points $\{v_1, \dots, v_n\}$ in $\mathbb{R}^d$ satisfies the ℓ_2^2 metric if for all i, j, k we have $|v_i - v_j|^2 + |v_j - v_k|^2 \geq |v_i - v_k|^2$.

An ℓ_2^2 representation of a graph is a mapping of vertices to points in some Euclidean space that satisfy the ℓ_2^2 metric. It is a *unit ℓ_2^2 representation* if points are of unit length.

Remark 2.3. By Pythagorean theorem, ℓ_2^2 requirement is equivalent to forcing the angle between $v_i - v_j$, $v_j - v_k$ to be at most $\pi/2$. See Figure 1.

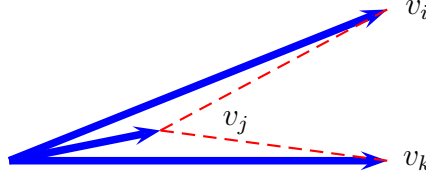


Figure 1: The blue are vectors v_i, v_j, v_k and in ℓ_2^2 metric we require that the angle formed at the red dashed lines is at most $\pi/2$.

Example 2.4. Suppose we have $n = 2^d$ vertices. Then we can map each vertex into the unit hypercube

$$\mathbf{HCUBE} = \frac{1}{\sqrt{d}} \{-1, 1\}^d$$

which forms a ℓ_2^2 metric (to see this, use the triangle inequality of the Hamming distance).

The second condition for the mapping that is needed is to require large separation in the distances.

Definition 2.5 (c -spread). An ℓ_2^2 representation i (vertex) $\rightarrow v_i \in \mathbb{R}^d$ is c -spread if $\sum_{i < j} |v_i - v_j|^2 \geq 4c(1 - c)n^2$.

Example 2.6. Map half of all vertices to $(-1, -1, -1, \dots)$ and other half to $(1, 1, 1, \dots)$. This is c -spread since sum of all distance is n^2 . However, $\mathbf{HCUBE}$ is NOT c -spread

$$\sum_{i < j} |v_i - v_j|^2 = \sum_{k=1}^d \frac{(2k)^2}{d} \binom{d}{k} = 2^{d+1} + (d-1)2^d = 2n + n(\lg n - 1)$$

Solving the following SDP ($\mathbf{P}$) will allow us to obtain a c -spread unit ℓ_2^2 representation with the optimal objective value being a lower bound on optimal solution to balanced cut.

$$(\mathbf{P}) \quad \min \quad \frac{1}{4} \sum_{ij \in E} |v_i - v_j|^2 \quad (1)$$

$$\forall i \quad |v_i| = 1 \quad (2)$$

$$\forall i, j, k \quad |v_i - v_j|^2 + |v_j - v_k|^2 \geq |v_i - v_k|^2 \quad (3)$$

$$\sum_{i < j} |v_i - v_j|^2 \geq 4c(1 - c)n^2 \quad (4)$$

Corollary 2.7. *The min value of (P) is a lower bound on the min c -balanced cut.*

Proof. Fix any unit vector u . Take an optimal solution $S^* \subseteq V$ for min c -balanced cut. Then map each $v \in S^*$ to u and each $v' \notin S^*$ to $-u$. The objective value is equal to $|\delta(S^*)|$, and it is easy to check the other conditions hold. Thus the defined mapping is a feasible solution to the SDP. $\square$

3 Main Result and Algorithm

Theorem 3.1 (Main). *There exists an algorithm (namely Algorithm 2 **FindLargeSets**) that finds two sets S, T each of size at least $c'n$ such that any pair $(v_i, v_j) \in S \times T$ has distance $|v_i - v_j| \geq \Delta = \Omega(\log^{-2/3} n)$*

Remark 3.2. In the ARV paper, they prove that a modified algorithm can output sets that have distance $\Omega(\sqrt{\log n})$. This yields a $\sqrt{\log n}$ approximation ratio. However, since that result relies heavily on the techniques for proving Theorem 3.1 (including a key technical lemma that will be proved here) we will not include the stronger results here.

With Theorem 3.1 it is possible to give a $\frac{1}{\Delta} = O(\log^{2/3} n)$ approximation algorithm as follows:

Algorithm 1: A randomized $O(\log^{2/3} n)$ approximation algorithm for c' -balanced cut. See Figure 2

BALANCEDCUT($G(V, E)$)

- (1) $\{v_i\} \leftarrow$ Solve SDP (P)
- (2) $(S, T) \leftarrow$ **FindLargeSets** ($v_1, \dots, v_n$)
- (3) $r \leftarrow \text{Uniform}[0, \Delta]$
- (4) Output $V_r \leftarrow$ Set of vertices v_i that has distance $d(v_i, S) \leq r$ (i.e. at most r to S).

Corollary 3.3. *Assuming that **FindLargeSets** produces Δ distant sets S and T of size at least $c'n$, **BalancedCut** finds a c' balanced cut $C \subseteq V$ such that $|\delta(C)| \leq O(\log^{2/3} n)OPT$ with high probability.*

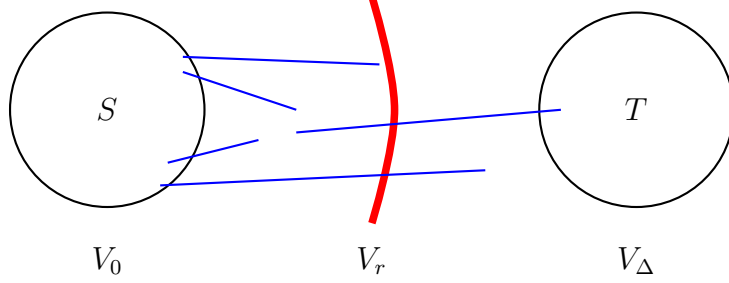


Figure 2: An illustration of **BalancedCut** algorithm. V_r are all vertices that have distance at most r to S . Blue lines are edges.

Algorithm 2: Finds sets each of size $c'n$ that are $\Omega(\log^{-2/3} n)$ distant. See Figure 3
FINDLARGESETS($v_1, \dots, v_n$ a c -spread unit ℓ_2^2 representation in $\mathbb{R}^d$)

- (1) **(STEP 1: Obtaining initial S and T)**
- (2) Pick unit vector $u \in \mathbb{R}^d$ uniformly at random.
- (3) $m \leftarrow \text{median}(\langle v_1, u \rangle, \langle v_2, u \rangle, \dots)$
- (4) $S \leftarrow \{v_i : \langle v_i, u \rangle \leq m - \frac{\sigma}{\sqrt{d}}\}$
- (5) $T \leftarrow \{v_i : \langle v_i, u \rangle \leq m + \frac{\sigma}{\sqrt{d}}\}$
- (6) **if** $|S| < 2c'n$ or $|T| < 2c'n$
- (7) **FAIL.**
- (8) **else**
- (9) **(STEP 2: Removing close points between S and T)**
- (10) **while** exists $(v_i, v_j) \in S \times T$ with $|v_i - v_j| < \Delta$
- (11) Remove v_i from S and v_j from T
- (12) **if** $|S| \geq c'n$ and $|T| \geq c'n$ **then return** S, T
- (13) **else FAIL.**

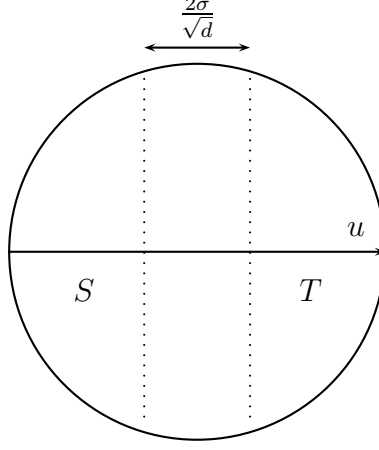


Figure 3: **FindLargeSets** takes points projected farther away from the median.

Proof. Let W be value of optimum solution to SDP $(\mathbf{P})$, and for $e = \{v_i, v_j\}$ let $w_e = \frac{1}{4}|v_i - v_j|^2$. We have $W \leq OPT$ from Corollary 2.7. First observe that **BalancedCut** produces a feasible solution. Let E' be edges from E removing edges where both endpoints are in S , or both endpoints are in T . For any $e = \{v_i, v_j\} \in E'$, the cut of $V_{r(e)}$ where $r(e) \in (d(v_i, S), d(v_j, S))$ contains e . The length of interval $(d(v_i, S), d(v_j, S))$ is less than w_e by triangle inequality. We have

$$\begin{aligned}
\mathbb{E}_{r \sim \text{Uniform}[0, \Delta]}[|\delta(V_r)|] &= \frac{1}{\Delta} \int_0^\Delta \sum_{e \in E'} \mathbf{1}(e \in \delta(V_r)) dr \\
&= \frac{1}{\Delta} \sum_{e = \{v_i, v_j\} \in E'} \int_0^\Delta \mathbf{1}(r \in (d(v_i, S), d(v_j, S))) dr \\
&\leq \frac{1}{\Delta} \sum_{e \in E'} w_e \leq \frac{W}{\Delta} \leq O(\log^{2/3} n) OPT
\end{aligned}$$

where $\mathbf{1}$ is the indicator function. Probability of success can be boosted. $\square$

It remains to show that **FindLargeSets** indeed finds large sets with large distance, this will occupy the remainder of these notes.

4 Proof of Main Theorem excluding Key Lemma

Proof outline:

1. We must first prove that Step 1 of **FindLargeSets** succeeds with high probability. This is where the c -spread condition is necessary, and some facts about random projections. See Section 4.1.

2. Proving step 2 succeeds with high probability is the main technical contribution of ARV.

The proof is roughly as follows: if we suppose step 2 fails with significant probability, then the deleted pairs forms a matching $M(u)$ for each of the large fraction of directions u . A key lemma we later develop tells us that for a large fraction of directions, there exists a pair (v_i, v_j) of distance at most k in the union of all $M(u)$'s, such that $|v_i - v_j|$ is relatively small compared to $\langle v_i - v_j, u \rangle$. And this cannot occur with significant probability for a suitable choice of k .

4.1 Properties of Projections onto Random Directions

Some properties about projecting fixed vectors onto random unit vectors will be useful for showing that the first step of **FindLargeSets** succeeds, and also in proving the key technical lemma.

Lemma 4.1. *Let $v \in \mathbb{R}^d$ then over random uniformly drawn unit vectors u ,*

1. $\mathbb{P}_u \left(|\langle v, u \rangle| \leq \frac{t|v|}{\sqrt{d}} \right) \leq 3t$, whenever $t \leq 1$
2. $\mathbb{P}_u \left(|\langle v, u \rangle| \geq \frac{t|v|}{\sqrt{d}} \right) \leq \exp(-t^2/4)$, whenever $t \leq \sqrt{d}/4$

Remark 4.2. These probabilities essentially say that (1) chances of the projection length being a factor t smaller is proportional to t and (2) there is very small chance that the projection is large. These are well known facts and are used in dimensionality reduction (for say clustering).

Proof. (Sketch, this is not in ARV) One way to see why it's true is to see that projection to a random direction is the same as fixing the direction and picking v randomly. Picking v randomly is same as picking it from a spherical Gaussian (with identity covariance matrix) and normalizing the length. Then projecting v to say $(1, 0, 0, \dots)$ is picking out the first coordinate of v which is generated from a univariate Gaussian – one can then apply Gaussian tail bounds. In fact, this is pretty much how one can prove the Johnson-Lindenstrauss theorem (with k random unit vectors). $\square$

To show that the first step of **FindLargeSets** succeeds with constant probability we show that the v_i 's satisfy the following definition.

Definition 4.3. $v_1, \dots, v_n$ is (σ, γ, c') *projection separated* if for at least γ fraction of directions u , there are two sets A, B of size at least $2c'n$ and

$$\begin{aligned} A &= \{v_i : \langle v_i, u \rangle \geq m + \sigma/\sqrt{d}\} \\ B &= \{v_i : \langle v_i, u \rangle \leq m - \sigma/\sqrt{d}\} \end{aligned}$$

where $m = \text{median}(\langle v_1, u \rangle, \dots, \langle v_n, u \rangle)$.

Lemma 4.4. *For every $c < 1/3$ there exists c', σ, γ such that any c -spread representation is (σ, γ, c') projection separated*

Remark 4.5. Notice that we do not require the ℓ_2^2 property.

Proof. (Sketch) The c -spread condition implies that $|v_i - v_j| \geq c(1 - c)/2$, and using property 1 about random projections, this means $|\langle v_i - v_j, u \rangle| \geq \Omega(c(1 - c)/\sqrt{d})$ with constant probability. Letting $\sigma = c(1 - c)$ we get that the number of pairs is larger than $\Omega(c(1 - c)n^2)$ for at least $\gamma = \Theta(c(1 - c))$ measure of directions.

See Lemma 6 of the paper for more details. $\square$

4.2 Covers and Matching Covers

We have already seen in Section 3 that step 1 of **FindLargeSets** succeeds with constant probability (of course we can boost this by running step 1 many times). What about step 2? We show later on that it cannot fail with probability $> \delta$. Then use union bound to get that **FindLargeSets** succeeds (in both stages) with high probability. Let us first define what must happen structurally (i.e. many pairs are deleted).

Definition 4.6. A (σ, δ, c') *matching cover* of $\{v_1, v_2, \dots, v_n\}$ is a collection $\mathcal{M}$ of matchings such that for $\geq \delta$ fraction of directions u , there exists a matching $M(u) \in \mathcal{M}$ of size at least $c'n$ where each pair $(v_i, v_j) \in M(u)$ has

$$\langle v_i - v_j, u \rangle \geq \frac{2\sigma}{\sqrt{d}}$$

The *associated matching graph* is the multigraph constructed from the union of all such $M(u)$'s.

Example 4.7. Consider the example in Figure 4 in $\mathbb{R}^2$ and the vectors a, b, c, d . Suppose that $u = (x, \sqrt{1 - x^2})$ (similar analysis holds when u dips below the x -axis), and that (a, b) and (c, d) are matched up. If $x \geq \sigma$,

$$\langle b - a, u \rangle = \langle (\sqrt{2}, 0), (x, \sqrt{1 - x^2}) \rangle = \sqrt{2}x \geq \frac{2\sigma}{\sqrt{2}}$$

What is the fraction of directions u for which its $x \geq \sigma$? When $x = \sigma$, the angle formed between u and x -axis is $\sin^{-1} \sigma$, so the fraction of directions for which the assigned matchings produce large projections is $\delta = \Theta(\frac{\sin^{-1} \sigma}{2\pi})$

Lemma 4.8 (Failure of step 2 implies matching covers). *If step 2 of **FindLargeSets** fails with probability $> \delta$, then there is a (σ, δ, c') matching cover.*

Proof. When step 2 fails for at least δ fraction of directions u , then we are left with a set of deleted pairs $M(u) = \{(v_i, v_j) : v_i \in S_u, v_j \in T_u\}$ of size at least a constant factor, c' of n , which we take as our matching for u . $\square$

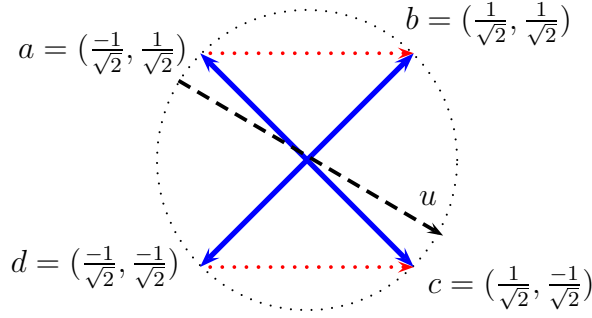


Figure 4: Points on the square shown in blue lines form a c -spread ℓ_2^2 representation. They also have a matching cover indicated by red lines.

It will be more useful to work with a similar definition that requires each v_i to be in some matching for a non-negligible fraction of the directions.

Definition 4.9. A collection of matchings $\mathcal{M}$ is a *uniform (σ, δ) matching cover* of a finite $X \subset \mathbb{R}^d$ if for every direction u , there exists $M(u) \in \mathcal{M}$ such that if $(v_i, v_j) \in M(u)$ then $|\langle v_i - v_j, u \rangle| \geq \frac{2\sigma}{\sqrt{d}}$. Further, every v_i is matched up in $M(u)$ for at least δ fraction of u 's.

Example 4.10. The example in Example 4.7 also happens to be a uniform (σ, δ) matching cover where each $v \in \{a, b, c, d\}$ participates in a matching in δ fraction of u 's. However, removing the edge (c, d) from the matching does not constitute a uniform matching because c, d do not participate with non-negligible probability.

Matching covers imply existence of uniform matching covers.

Lemma 4.11. Given n points that are (σ, δ, c') matching covered by $\mathcal{M}$, then there exists a subset X of size $\Omega(c'\delta n)$ such that they are uniformly $(\sigma, \Omega(\delta c'))$ matching covered by $\mathcal{M}$.

Proof. See paper for details. □

The following definition of a covering will be needed in proving the key lemma.

Definition 4.12 (Cover). Given a finite set $Z = \{z_1, z_2, \dots, z_m\} \subseteq \mathbb{R}^d$ with $|z_i| \leq \ell$, it is an (ϵ, δ) -cover if there are δ fraction of directions such that for each direction u , there exist $z_i \in Z$ with $\langle z_i, u \rangle \geq \epsilon$. The δ is called the *covering probability* and ϵ is the *projection length*.

Definition 4.13 (Cover of a Point). A point x is (ϵ, δ) -covered by Z if the set $\{z - x : z \in Z\}$ is an (ϵ, δ) -cover.

Remark 4.14. Covering is a much weaker notion than matching covers. Intuitively, one can think of directions u participating in a cover as simply the points on the unit sphere with

close distance to any point in the cover (hence, large projections). Each direction u requires exactly one “close” vector in the cover.

However, a matching cover requires, for each u , that a *constant fraction* of the points match up and the difference of *all pairs* have large projection on u .

To illustrate, the points $\mathbf{HCUBE} = \{1/\sqrt{d}, -1/\sqrt{d}\}^d$ has a cover with a constant covering probability. To see this, note that the covering probability is equal to the mass of all points within a certain distance to any point in the cover, then apply the measure concentration inequality in Lemma 5.1. However, as stated in the ARV paper, $\mathbf{HCUBE}$ only has a $(\Theta(1), \Theta(1/\sqrt{d}))$ matching cover.

4.3 Key Lemma and Proof of Main Theorem

In this section we will present the key technical lemma and why it implies `FindLargeSets` succeeds with good probability. First, let us define a set of points where each point is an endpoint for large projections to random directions, for large fraction of directions.

Definition 4.15. Given n points that is (σ, δ, c') matching covered by $\mathcal{M}$, define $v \in S_{k, \xi}$, the (k, ξ) *core*, if v is $(k\sigma/(2\sqrt{d}), \xi)$ covered by vectors that are within distance k of v in the associated matching graph of $\mathcal{M}$.

For convention, let $S_k = S_{k, 1/2}$. Now we are ready to state the key technical lemma that asserts either the existence of large k -cores or two vectors whose difference has large length.

Lemma 4.16 (Key Lemma). *Given a set of n vectors that is (σ, δ, c') matching covered, there exists constants α and g depending on δ, c' such that for all $k \geq 1$, either one of the following must occur*

1. $|S_k| \geq \alpha^k n$.
2. There exists v_i, v_j with distance k in the matching graph such that $|v_i - v_j| \geq g\sigma/\sqrt{k}$.

We delay its proof into Section 5. But now we show how this implies good probability of success of step two for `FindLargeSets`.

Theorem 4.17 (Main). *`FindLargeSets` finds, with probability at least δ , $\Omega(\log^{-2/3} n)$ separated sets given a unit ℓ_2^2 representation.*

Proof. By Lemma 4.8, if `FindLargeSets` fails with probability at least $1 - \delta$ then there exists a (σ, δ, c') matching cover. Every edge (v_i, v_j) in the matching graph for some direction u has $\langle v_i - v_j, u \rangle \geq 2\sigma/\sqrt{d}$, yet $|v_i - v_j| < \sqrt{\Delta}$.

Since we have a matching cover we can apply the Key Lemma to show that neither conditions are met, and therefore arrive at a contradiction and conclude that `FindLargeSets` should succeed with probability at least δ . Let us choose $k = g\sigma/\sqrt{\Delta}$, where g is from the lemma, to proceed.

Is the second condition satisfied? For any v_i, v_j within distance k in the matching graph for the matching cover we have, from the triangle inequality for ℓ_2^2 metric, $|v_i - v_j|^2 \leq k\Delta$ so that $|v_i - v_j| \leq g\sigma/\sqrt{k}$ when we sub in Δ in terms of k . So this condition is not satisfied.

Is the first condition satisfied? Recall that this condition requires for at least half of all directions u , there exists v_i, v_j such that $|v_i - v_j| \leq \sqrt{k\Delta}$ (they are k distant in the matching graph), and they've got large projection to u : $\langle v_i - v_j, u \rangle \geq k\sigma/(2\sqrt{d})$.

Intuitively this is unlikely to happen, you cannot have the best of both worlds. Formally, use the property about Gaussian projections. That is, what is the probability that the projection of $v_i - v_j$ to u is t times larger than its length $\sqrt{k\Delta}$ over $\sqrt{d}$? Where $t = \Theta(\sqrt{k}\sigma/\sqrt{\Delta})$ so the projection is at least as stated above. It is less than $\exp(-\Theta(t^2)) = \exp(-\Omega(\log n)) = 1/n^{\Theta(1)}$. If we choose the constants properly, we can reduce the probability to no more than $\frac{1}{2n^2}$. Then we simply union bound the probability of this happening to any pair of points, which is $\frac{n^2}{2n^2} = 1/2$ (since there's $< n^2$ possible pairs). The probabilities are over random u 's so this contradicts the first condition. $\square$

5 Proof of Key Lemma

5.1 Concentration of Measure

We first need a useful tool that will allow us to boost the fraction of directions that is covered for a cover of a point by slightly decreasing the projection length required. Define S^{d-1} to be the unit sphere in $\mathbb{R}^d$ and μ the standard measure on it. For $A \subseteq \mathbb{R}^d$, let $A_\gamma = \{x \in \mathbb{R}^d : \exists y \in A, |x - y| \leq \gamma\}$. The boosting result will follow directly from the following

Lemma 5.1 (Concentration of Measure). *If $t > 0$, $A \subseteq S^{d-1}$ is measurable and $\gamma > \frac{2\sqrt{\log(1/\mu(A))+t}}{\sqrt{d}}$ then $\mu(A_\gamma) \geq 1 - \exp(-t^2/2)$.*

Proof. The proof will not be covered as it relies on the non-trivial Levy's isoperimetric inequality [2]. $\square$

Lemma 5.2 (Boosting Covering Probability). *Let $C = \{v_1, v_2, \dots\} \subseteq \mathbb{R}^d$ be an (ϵ, δ) -cover with $|v_i| \leq \ell$. We have for any $\gamma > \frac{2\sqrt{\log(1/\mu(A))+t}}{\sqrt{d}}$, C is a $(\epsilon - 2\ell\gamma, 1 - \exp(-t^2/2))$ -cover.*

Proof. Let A be the set of unit vectors u that have projection length $\langle u, v_i \rangle \geq \epsilon$. Thus by definition of C , $\mu(A) \geq \delta$. Invoking Lemma 5.1 with assumption on γ , $\mu(A_\gamma) \geq 1 - \exp(-t^2/2)$.

It remains to show $u' \in A_\gamma$ has projection length $\geq \epsilon - 2\ell\gamma$ on C . Take the closest $u \in A$ to u' so that $|u - u'| \leq \gamma$ and let $v_i \in C$ be the closest to u in projection. Let θ be the angle between v_i and u (which must be non-obtuse), and θ' the angle between v_i and u' .

What is the angle α between u and u' ? Consider the isosceles triangle formed with sides u and u' having length 1, and $u - u'$ has length at most γ . Thus we have $\gamma/2 \geq \sin(\alpha/2)$,

applying the identity $2 \sin x \geq x$ we get $\gamma \geq \alpha/2$ so angle is at most 2γ . The angle θ' is maximized when v_i, u, u' lie on a plane with u in the middle, taking note of the previous fact, we get that $\theta + 2\gamma \geq \theta'$. We can write $\langle v_i, u' \rangle = |v_i| \cos \theta'$ and note the same for u . We get

$$|v_i| \cos \theta' \geq |v_i| \cos(\theta + 2\gamma) \geq |v_i| \cos \theta - 2|v_i|\gamma \geq \epsilon - 2\ell\gamma$$

where the 2nd last inequality comes from the fact that $\cos$ has maximum absolute slope of 1. $\square$

5.2 Cover Composition

Now we show a construction where given a set of covered points D and a uniform matching X , we can construct a set of covered points D' each having projection length that of D plus that of X . This will prove useful for constructing a k -core that is central to the key lemma.

Lemma 5.3 (Cover Composition). *Suppose $X \subseteq \mathbb{R}^d$ is uniform (ϵ_X, δ) -covered, and $D \subseteq X$ where every point $v_i \in D$ is $(\epsilon_D, 1 - \delta/2)$ covered by D_i .*

If $|D| \geq \tau|X|$ then there exists a set $D' \subseteq X$ with $|D'| \geq \tau\delta|X|/4$ that are $(\epsilon_X + \epsilon_D, \tau\delta/4)$ covered. For each $v_k \in D'$ the vectors of its cover have the form

$$v_j - v_k = v_j - v_i + v_i - v_k$$

where $v_j - v_i \in D_i$ and (v_i, v_k) is in the matching graph of X (i.e. they are matched up for some direction).

Proof. Fix $v_i \in Z$, then for each of the $1 - \delta/2$ fraction of directions u , there exists $v_j \in D_i$ in its cover such that $\langle v_j - v_i, u \rangle \geq \epsilon_D$. For each of the δ fraction of directions u , the uniform matching tells us that there is a point v_k with $\langle v_i - v_k, u \rangle \geq \epsilon_X$ (i.e. v_k is matched up with v_i in the matching graph for X).

What fraction of directions do both events happen? At most $\delta/2$ fraction of directions are “bad” for the former event, but at least δ fraction are “good” for the latter event, so $\delta - \delta/2 = \delta/2$ fraction are “good” for both events. Adding up the dot products in both cases, $\langle v_j - v_k, u \rangle \geq \epsilon_D + \epsilon_X$.

Note that $k = k(i, u)$ since there is exactly one match for v_i in the matching graph for a fixed direction. Let us also say $v_j - v_k$ is “assigned” to vector v_k . See Figure 5 that shows this composition.

Since $\delta/2$ fraction of directions have $v_j - v_k$ assigned to v_k , computing the expectation over uniform random choice of direction u , it has at least $\delta|D|/2$ vectors (since we can sum over i of $k(i, u)$). Since an average of number of assigned vectors for u equals $|X|$ times probability p of a random vector being assigned for u we have $p = \delta|D|/(2|X|)$ which is also the expected mass of directions where a (randomly chosen) vector gets assigned.

Noting the fact that $k = k(i, u)$, we get that at least $\delta|Z|/4 \geq \delta\tau|X|/4$ points in X have assigned vectors for at least $\delta|D|/(4|X|) = \delta\tau/4$ mass of the directions. This set of points is our D' . $\square$

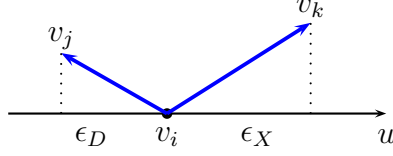


Figure 5: v_j gets assigned to the cover of v_k , and $v_j - v_k$'s projection on u is composed from that of $v_j - v_i$ and $v_i - v_k$.

Remark 5.4. It is possible to boost the D' of covered points to probability $1 - \delta/2$ while reducing the projection length by $\epsilon_X/2$. This can be seen by applying Lemma 5.2, setting $1 - \delta/2 = 1 - \exp(-t^2/2)$ solving and getting $t = \log(2/\delta)$. Now, setting $2\ell\gamma = \epsilon_X/2$ (plugging in t as well), we solve for ℓ and get

Corollary 5.5. *If vectors in the cover of each point of D' have length at most*

$$\frac{\epsilon_X \sqrt{d}}{8\sqrt{\log \frac{8}{\tau\delta}} + 4\sqrt{\log \frac{2}{\delta}}}$$

then $(\epsilon_X + \epsilon_D, \tau\delta/4)$ covers of D' are also $(\epsilon_X/2 + \epsilon_D, 1 - \delta/2)$ covers.

5.3 Bringing it Together

Before carrying on, note that the condition in the lemma of requiring a set of points that is (σ, δ, c') matching covered implies, via Lemma 4.11 that there exists a subset X that is uniform (σ, δ) matching covered. We are now ready to prove a stronger version of the key lemma:

Lemma 5.6. *Fix δ , then for all $k \geq 1$, we have either*

1. $|S_{k, 1-\delta/2}| \geq (\delta/4)^k |X|$
2. *There exists v_i, v_j with $|v_i - v_j| \in \Omega(\sigma/\sqrt{k})$, and they are distance at most k in the associated matching graph for X .*

Proof. Proceed by induction on k . When $k = 1$, we note that X is uniformly (σ, δ) matching covered, which means that each $v \in X$ is matched up with v' in at least δ fraction of directions, and for each direction u , this means $\langle u, v - v' \rangle \geq \sigma/(2\sqrt{d})$. In other words, $X = S_{1, \delta/2}$.

Supposing the lemma holds for all integers $\leq k$, we first note that if the latter case holds for k , then use the same pair of vectors so it holds for $k + 1$ (the length of their difference and their distance in matching graph are the same).

What about former case? We can apply the cover composition lemma (Lemma 5.3) on $S_{k,\delta/2}$ (since every point is covered there) and on the uniform matching cover for X to obtain S' of size $\geq (\delta/4)|S_{k,\delta/2}| \geq (\delta/4)^{k+1}|X|$, that has projection length

$$\epsilon_X + \epsilon_{S_{k,\delta/2}} = \frac{2\sigma}{\sqrt{d}} + k \frac{\sigma}{2\sqrt{d}} \geq (k+1) \frac{\sigma}{2\sqrt{d}}$$

and covering probability $(\delta/4)^{k+1}$. The cover composition construction ensures that the vectors in the covers of S are $k+1$ distance away from the point that they cover. We need to boost the probability to $1 - \delta/2$. Applying Corollary 5.5 with $\tau = (\delta/4)^k$ we get that if the length of the covers of S' are at most

$$\ell \leq \frac{2\sigma}{8\sqrt{(k+1) \ln \frac{O(1)}{\delta}} + 4\sqrt{\ln \frac{2}{\delta}}}$$

then the decrease in projection length is at most $\epsilon_X/2 = \sigma/\sqrt{d}$ so we can let $S_{k+1,\delta/2} = S'$ and the first constraint is met. Otherwise there is some vector of the form $v_i - v_j$ in some cover of some point in S' with length $\Omega(\sigma/\sqrt{k+1})$ so the second condition is met. $\square$

References

- [1] S. Arora, S. Rao, and U. V. Vazirani. Expander flows, geometric embeddings and graph partitioning. In L. Babai, editor, *STOC*, pages 222–231. ACM, 2004.
- [2] K. Ball. *Flavors of Geometry*, chapter An elementary introduction to modern convex geometry. Cambridge University Press, 1997.
- [3] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J. ACM*, 46(6):787–832, 1999.