MCS 549 Project: Bilu-Linial stability [1]

Gregoire Fournier

March 3, 2025

1 Intro

Informally, an instance of a problem is Bilu-Linial stable if the optimal solution does not change when the instance is perturbed.
Such instances yields more robust solutions than worst cases in many problems, and are closer to real word application.
Some studied applications for a perturbation α are:

- Graph partitioning, $G = (V, E, w)$ and $G' = (V, E, w)$ st. $\forall e \in E, w(e) \leq w'(e) \leq \alpha w(e)$.
- Clustering $J = (V, d)$ and $J' = (V, d')$ st. $\forall u, v \in V, d(u, v) \leq d'(u, v) \leq \alpha d(u, v)$.

An instance is α-stable if its every α perturbation does not change the optimal solution.
When studying LPs-SDPs, an instance is α-weak Bilu-Linial stable for an (α, ε) perturbation resilience.

problem	**main results**	**reference**
Max Cut & 2-correlation clustering	$O(\sqrt{\log n} \log \log n)$ (incl. weakly stable instances) SDP gap and hardness result	Makarychev et al. (2014b)
Min Multiway Cut	4, (incl. weakly stable instances)	Makarychev et al. (2014b)
Max k-Cut	hardness for ∞-stable instances	Makarychev et al. (2014b)
sym./asym. k-center	2 hardness for $(2-\varepsilon)$-pert. resil.	Balcan et al. (2015)
s.c.b. objective	$1+\sqrt{2}$ $(2+\sqrt{3}, \varepsilon)$ for k-median 2, assuming cluster verifiability	Balcan and Liang (2016) Balcan et al. (2015)
s.c.b., Steiner points	$2+\sqrt{3}$	Awasthi et al. (2012)
min-sum objective	$O(\rho)$ and $(O(\rho), \varepsilon)$, where ρ is the ratio between the sizes of the largest and smallest clusters	Balcan and Liang (2016)
TSP	1.8	Mihalák et al. (2011)

Figure 1: Known results for Bilu-Linial stability [1]

2 Clustering: Stable instances

2.a Metric perturbation resilience for k-Center [2]

Motivation: Fast algorithm to solve 2-metric perturbation resilient instances of k-center.

Note that given a set of centers or a clustering, it is possible to efficiently find the corresponding clustering or the corresponding optimal set of centers.

Definition 1 (k-Center) Given vertices V, a metric d, in V define k-centers, $c_1, ...c_k$ which induce a clustering $C_1, ...C_k$ on V based on the d-nearest centers. Hence:

$$C_i = \{u : \forall j \neq i, d(u, c_i) \leq d(u, c_j)\}$$

$$cost = max_i max_{u \in C_i} d(u, c_i)$$

Theorem 2 (Stability and approximation for k-center) Every α-approximation algorithm finds the optimal solutions of α-metric perturbation resilient instances.

Proof. Consider $(C_i)_i$ the optimal clustering solution with cost r^* and $(C_i')_i$ the approximation. By definition of an approximation $\forall i, \forall u \in C_i^{'}, d(u, c_i') \leq \alpha r^*$. Now define d':

$$\forall u, v \in V, d'(u, v) = \begin{cases} d(u,v)/\alpha & \text{if } d(u,v) \geq \alpha r^* \\ r^* & \text{if } d(u,v) \in [r^*, \alpha r^*] \\ d(u,v) & \text{if } d(u,v) \leq r^* \end{cases}$$

So that d' defines a distance, it must satisfy the triangle inequality, so define f:

$$f(x) = \begin{cases} 1/\alpha & \text{if } x \geq \alpha r^* \\ r^*/x & \text{if } x \in [r^*, \alpha r^*] \\ 1 & \text{if } x \leq r^* \end{cases}$$

Then $d' = (f \circ d)d$. Note that f is non increasing, $xf(x)$ non decreasing. Now for u, v, w, suppose wlog $d(u, w) \geq max(d'(u, v), d'(v, w))$, want to prove $d'(u, v) + d'(v, w) \geq d(u, w)$. Since $xf(x)$ non decreasing $f(d(u, w)) \geq min(f(d(u, v)), f(d(v, w)))$ and then:

$$d'(u, v) + d'(v, w) = ((f \circ d)d)(u, v) + ((f \circ d)d)(v, w) \geq (f \circ d)(u, w)(d(u, v) + d(v, w))$$

So by triangular inequality in d:

$$d'(u, v) + d'(v, w) \geq (f \circ d)(u, w)d(u, w) \geq d'(u, w)$$

d' is an α perturbation as $\forall u, v \in V, \frac{d'(u,v)}{d(u,v)} = f(d(u, v)) \in [1/\alpha, 1]$. Therefore $C_1, ..., C_k$ is still the optimal clustering for d' by definition of α resilience. Denote $c_1", ..., c_k"$ the optimal set of centers for d' of $C_1, ..C_k$.

Define $r(C_i) = min_{c \in C_i} max_{u \in C_i} d(u, c)$ and fix i st $r(C_i) = r^*$. And in particular, $\exists u, c \in C_i$ st $d(u, c) \geq r(C_i) = r^*$ and then $d(u, c_i") \geq r^*$ as well. So the cost of $C_1', ...C_k'$ is at least r^*, and cannot be more than r^* since $d' \leq d$ on V. So the cost of the clustering $C_1', ..C_k'$ of V

according to d' is r^*.

Finally the cost of $C'_1, ..C'_k$ is also r^*: Denote $c'_1, ..., c'_k$ the optimal set of centers for d',as $\forall i, \forall u \in C'_i, d(u, c'_i) \leq \alpha r^*$ and so $d'(u, c'_i) \leq r^*$. Therefore $C'_1, ..C'_k$ is an optimal clustering for d', and it must be equal to $C_1, ..C_k$. ■

2.b Clustering problems with separable center-based objectives [3]

Definition 3 (Separable center-based objectives) A clustering problem has a center-based objective if the following holds:

- Given $S \subset V$ and a distance d_S on S, it is possible to find the optimal center or set of optimal centers (subset of S).
- The set of centers does not change when multiplying all distances in S by α.
- If $C_1, ..., C_k$ is an optimal clustering of V. Then $\forall i, \forall p \in C_i \ d(p, c_i) < d(p, c_j)$.

It is separable if:

- The cost of the clustering is either the maximum or sum of the cluster scores.
- The score $score(S, d|S)$ of each cluster S depends only on $(S, d|S$, and can be computed in poly time.

Many standard clustering problems, such as k-center, k-means have separable center-based objectives.
Now consider α-metric perturbation resilient instances ($\alpha = 1 + \sqrt{2}$).

Theorem 4 (α-center proximity property)

$$i \neq j, \forall p \in C_i, \ d(p, c_j) > \alpha d(p, c_i)$$

Proof. Otherwise $d(p, c_j) \leq \alpha d(p, c_i)$. Fix $r^* = d(p, c_i)$ and define d' as follows:

$$\forall u, v, \ d'(u, v) = min[d(u, v), d(u, p) + r^* + d(c_j, v), d(v, p) + r^* + d(c_j, u)]$$

The metric $d'(u, v)$ is the shortest path metric on the complete graph on V with edge lengths $len(u, v) = d(u, v)$ for all (u, v) except the edge (p, c_j) ($len(p, c_j) = r^*$).
Since $d(u, v)/len(u, v)$ is at most $d(p, c_j)/r^* \leq \alpha$ for all edges (u, v), then $\forall u, v, d(u, v)/d'(u, v) \leq \alpha$.
Therefore d' is an α-metric perturbation of d. Now to finish, show that $d' = d$ within C_i and within C_j. Then it yields:

$$d(c_i, p) = d'(c_i, p) < d'(cj, p) = r^* = d(c_i, p)$$

Which is a contradiction. ■

Proof. $d' = d$ within C_i and within C_j

Consider $u, v \in C_i$, to show $d(u, v) = d'(u, v)$, prove $d(u, v) \leq min(d(u, p)+r^*+d(cj, v), d(v, p)+r^* + d(cj, u))$. Wlog assume $d(u, p) + r^* + d(c_j, v) \leq d(v, p) + r^* + d(cj, u)$.
Then $d(u, p) + r^* + d(c_j, v) = d(u, p) + d(p, c_i) + d(c_j, v) \leq d(u, c_i) + d(c_j, v)$. If $v \in C_i$,

$d(v, c_i) < d(v, c_j)$, and thus $d(u, p) + r^* + d(c_j, v) > d(u, c_i) + d(c_i, v) \geq d(u, v)$.

Consider $u, v \in C_j$. Similarly to the previous case, need to show $d(u, v) \leq min(d(u, p) + r^* + d(c_j, v), d(v, p) + r^* + d(c_j, u))$. Now $u \in C_j$ so $d(u, c_j) < d(u, c_i)$. Thus, $d(u, p) + r^* + d(c_j, v) = (d(u, p) + d(p, c_i)) + d(c_j, v) \geq d(u, c_i) + d(c_j, v) > d(u, c_j) + d(c_j, v) \geq d(u, v)$.

■

To get to the algorithm quicker, some theorems are admitted:

Theorem 5 (Some properties from the previous results)

- All points outside of C_i lie at distance greater than r_i from c_i.
  So $C_i = B(c_i, r_i)$.
- Each $p \in C_i$ is closer to c_i than to any point q outside of C_i. Furthermore, $\forall p \in C_i$ and $q \notin C_i$, $\sqrt{2}d(p, c_i) < d(p, q)$.
- For $C_i \neq C_j$, $d(c_i, c_j) > \sqrt{2}max(r_i, r_j)$.

2.c A clustering algorithm in $O(nk)$

- Similar to building the minimum spanning tree, start by assigning a cluster to each vertex. For $n - 1$ steps, fusion the two nearest clusters. Assign a binary decomposition tree $\mathcal{T}$ to this process.
- Using a dynamic program algorithm, identify the best k C_is in $\mathcal{T}$.

The distance used is called the closure distance:

Definition 6 (Closure distance D_S)

$$\forall A_1, A_2 \subset V,\ D_S(A_1, A_2) \text{ is the minimum } r \text{ st. } A_1 \cup A_2 \text{ has an } r\text{-central point.}$$

$x \in A$ is r-central for A if:

- $A \subset B(x, r)$
- if $d(p, q) \leq d(p, x) \leq r$ then $d(q, x) \leq r$

Now the goal is to prove:

Theorem 7 (The C_is appear in $\mathcal{T}$) Consider a cluster C_i in the optimal clustering.

- Let C be a cluster/node in the decomposition tree. Then:

$$C \subset C_i,\ C_i \subset C \ \text{ or } \ C \cap C_i = \emptyset$$

- C_i appears in the decomposition tree.

References

[1] K. Makarychev. Y. Makarychev. Bilu–Linial Stability.

[2] Maria-Florina Balcan, Nika Haghtalab, and Colin White. Symmetric and asymmetric k-center clustering under stability. CoRR, abs/1505.03924, 2015.

[3] Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. CoRR, abs/1112.0826, 2011.