THE UNIVERSITY OF CHICAGO

CORRELATION CLUSTERING WITH LOCAL AND GLOBAL OBJECTIVES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY

JAFAR JAFAROV

CHICAGO, ILLINOIS

JUNE 2022

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

ABSTRACT

CHAPTER 1

INTRODUCTION

Grouping objects based on the similarity between them is a ubiquitous and important task in machine learning. This similarity information between objects can be represented in many ways, some of them being pairwise distances between objects (objects which are closer are more similar) or the degree of similarity between pairs of objects (objects which are more similar have a higher degree of similarity). Bansal, Blum, and Chawla [7] introduced the Correlation Clustering problem, a versatile model that elegantly captures this task of grouping objects based on similarity information. Since its introduction, the correlation clustering problem has found use in a variety of applications, such as co-reference resolution (see e.g., Cohen and Richman [21, 22]), cross-lingual link detection (see e.g., Van Gael and Zhu [46]), spam detection (see e.g., Ramachandran et al. [41], Bonchi et al. [14]), image segmentation (see e.g., Wirth [47]), multi-person tracking (see e.g., Tang et al. [44, 45]), computational biology (see e.g., Ben-Dor et al. [9]) and data mining (see e.g., Filkov and Skiena [28]).

In the Correlation Clustering problem, we are given a set of objects with pairwise similarity information. Our goal is to partition the objects into clusters that agree with this information *as much as possible*. Correlation Clustering is different from other classical clustering problems in that the given data is qualitative rather than quantitative. The pairwise similarity information is represented as a weighted graph G whose edges are labelled as "positive/similar" and "negative/dissimilar" by a noisy binary classifier. This stands in striking contrast with well-studied clustering problems such as k-means or k-median where objects are embedded in a metric space and a similarity information between two objects is given by a distance function. Furthermore, in Correlation Clustering the number of clusters is not given as a separate parameter; it solely depends on the instance and can vary between one and the number of vertices. Correlation Clustering can be seen as an agnostic learning problem: the goal is to find a classifier from a given hypothesis class (set of clusterings) that

fits to noisy examples (pairwise similarity information) as much as possible.

In Correlation Clustering the goal is to find a clustering $\mathcal{C}$ that is maximally consistent with the edge labels. A positive edge is in disagreement with $\mathcal{C}$, if its endpoints belong to different clusters; and a negative edge is in disagreement with $\mathcal{C}$ if its endpoints belong to the same cluster. The goal is to find a clustering $\mathcal{C}$ which minimizes the weight of edges in disagreement. We call this objective the MinDisagree objective.

Observe that if a binary classifier is not noisy, i.e., if there exists a clustering which is consistent with all the edge labels, then it is easy to find one: simply remove all negative/dissimilar edges and output the connected components of the remaining graph. Thus, the interesting case is when the binary classifier is noisy.

The MinDisagree objective has been extensively studied in the literature since its introduction by Bansal et al. [7]. There are currently two standard settings for Correlation Clustering which we will refer to as (1) Correlation Clustering on Complete Graphs and (2) Correlation Clustering with Noisy Partial Information. In the former setting, we assume that graph G is complete and all edge weights are the same, i.e., G is unweighted. In the latter setting, we do not make any assumptions on the graph G. Thus, edges can have arbitrary weights and some edges may be missing. These settings are quite different from the computational perspective.

For the Correlation Clustering on Complete Graphs model the first constant-factor approximation algorithm was given by Bansal et al. [7]. Charikar, Guruswami, and Wirth [16] gave a $4-$approximation algorithm for the problem. Ailon, Charikar, and Newman [3] gave two different algorithms with approximation factors of 3 and 2.5. Finally, a $2.06-$approximation algorithm was given by Chawla, Makarychev, Schramm, and Yaroslavtsev [19]. This is currently the best approximation guarantee for the MinDisagree objective. Some of these results (Charikar et al. [16], Ailon et al. [3], Chawla et al. [19]) rely on linear programming (LP) relaxation. Charikar et al. [16] gave an almost matching integrality gap instance with the integrality ratio of 2. Furthermore, Charikar et al. [16] showed that Correlation Clustering

Model	**Complete Unweighted Graphs**	**Arbitrary Weighted Graphs**
Approx. Factor	$\approx 2 \cdot 10^4$ [7] 4 [16] 3 and 2.5 [3] 2.06 [19]	$O(\log n)$ [16, 25]
Integrality Gap	2	$\Omega(\log n)$ [16, 25]
Hardness Result	APX-hard [16]	Equivalent to Multicut [16, 25]

Figure 1.1: Known Results for MinDisagree

on Complete Graphs is APX-hard, i.e., it is NP-hard to approximate the problem within a certain constant $c > 1$.

For the Correlation Clustering with Noisy Partial Information model Charikar et al. [16] and Demaine, Emanuel, Fiat, and Immorlica [25] independently gave an $O(\log n)$ approximation algorithm. Both algorithms use a linear programming relaxation and are heavily inspired by the region growing approach for the Multicut problem by Garg, Vazirani, and Yannakakis [29]. Both Charikar et al. [16] and Demaine et al. [25] complement their results with a matching integrality gap of $\Omega(\log n)$, thus implying that one cannot hope for a better approximation guarantee using an LP-based approach. Interestingly, Charikar et al. [16] and Demaine et al. [25] showed that this inspiration by Garg et al. [29] was not coincidental: they proved that in this setting the MinDisagree objective is equivalent to Multicut from the approximation perspective. In particular, they gave an approximation preserving reduction for both directions between two problems. This combined with the hardness result for Multicut by Chawla, Krauthgamer, Kumar, Rabani, and Sivakumar [18] show that $O(\log n)$ is likely to be the best possible approximation for this problem. More specifically, it is NP-hard to obtain any constant factor approximation algorithm for MinDisagree if the Unique Games Conjecture is true. See Table 1.1 for the summary of the results.

1.1 Correlation Clustering in Other Settings

Correlation Clustering has been considered in various settings. Chierichetti, Dalvi, and Kumar [20], Cohen-Addad, Lattanzi, Mitrovic, Norouzi-Fard, Parotsidis, and Tarnawski [23] studied the problem for the massively parallel computation model. Ahn, Cormode, Guha, McGregor, and Wirth [2] explored Correlation Clustering for the streaming model. Cohen-Addad, Fan, Lattanzi, Mitrovic, Norouzi-Fard, Parotsidis, and Tarnawski [24] considered it in the differential privacy setting. Mathieu, Sankur, and Schudy [38], and Lattanzi, Moseley, Vassilvitskii, Wang, and Zhou [34] studied the online version of Correlation Clustering.

Correlation Clustering on complete bipartite graphs was first studied by Amit [6], who presented an $11-$approximation algorithm. This was improved to a $4-$approximation algorithm by Ailon, Avigdor-Elgrabli, Liberty, and van Zuylen [4]. Chawla et al. [19] gave an algorithm with a $3-$approximation guarantee for complete $k-$partite graphs.

For the Correlation Clustering on Complete Graphs model Giotis and Guruswami [30] considered the case when the number of clusters is bounded by some constant. Under this assumption they gave a PTAS for MinDisagree.

Another interesting objective for Correlation Clustering is maximizing agreements. A positive edge is in agreement with $\mathcal{C}$, if its endpoints belong to the same cluster; and a negative edge is in agreement with $\mathcal{C}$ if its endpoints belong to different clusters. The goal is to find a clustering $\mathcal{C}$ which maximizes the weight of edges in agreement. We call this objective the MaxAgree objective. MaxAgree and MinDisagree are equivalent at optimality, but, differ from the approximation perspective. More specifically, the MaxAgree objective is easier than the MinDisagree objective. There is a trivial $2-$approximation algorithm for MaxAgree: if the weight of positive edges is greater than the weight of negative edges output G as a single cluster, otherwise output each vertex as a singleton cluster.

For the Correlation Clustering on Complete Graphs model Bansal et al. [7] gave a polynomial time approximation scheme (PTAS) for MaxAgree, i.e., it can be approximated

within any constant in polynomial time.

For the Correlation Clustering with Noisy Partial Information Charikar et al. [16] and Swamy [42] gave a $0.766-$approximation algorithm based on semidefinite programming (SDP) relaxation. Furthermore, Charikar et al. [16] and Tan [43] proved that MaxAgree is APX-hard for this setting.

Mathieu and Schudy [37] introduced a semi-random model for Correlation Clustering which can be defined as follows: given a ground truth clustering, $\mathcal{C}^*$, the classifier labels each edge to be consistent with $\mathcal{C}^*$ independently with probability p, otherwise decides for its label arbitrarily. Mathieu and Schudy [37] gave a $(1 + o(1))-$approximation. Their result was extended to other semi-random models by Makarychev, Makarychev, and Vijayaraghavan [35]. For a more detailed discussion see Section 2.1.

1.2 Correlation Clustering with the ℓ_p Objective

A natural generalization of MinDisagree is the ℓ_p objective. Define the disagreements vector to be a vector indexed by the vertices of G. Given a clustering $\mathcal{C}$, $\text{dis}(\mathcal{C}, E^+, E^-) \in \mathbb{R}^V$ is a $|V|$-dimensional vector where the u-th coordinate is equal to the weight of disagreements at u with respect to $\mathcal{C}$. That is,

$$\text{dis}_u(\mathcal{C}, E^+, E^-) = \sum_{(u,v)\in E} \mathbb{1}\{(u, v) \text{ is in disagreement with } \mathcal{C}\}$$

Thus, the MinDisagree objective is equivalent to finding a clustering minimizing the ℓ_1 norm of the disagreements vector. Another objective for Correlation Clustering that has received attention recently is to minimize the weight of disagreements at the vertex that is worst off (also known as Min Max Correlation Clustering). This is equivalent to finding a clustering that minimizes the ℓ_∞ norm of the disagreements vector. Observe that minimizing the ℓ_1 norm is a global objective since the focus is on minimizing the total weight of disagreements.

In contrast, for higher values of p (particularly $p = \infty$), minimizing the ℓ_p norm becomes a more local objective since the focus shifts towards minimizing the weight of disagreements at a single vertex. Minimizing the ℓ_2 norm of the disagreements vector can thus provide a balance between these global and local perspectives – it considers the weight of disagreements at all vertices but penalizes vertices that are worse off more heavily. The following scenario is a showcase that minimizing the ℓ_2 norm might be a more suitable objective than minimizing the ℓ_1 norm. Consider a recommender system such that input is a bipartite graph with left and right sides representing customers and services, respectively. A positive edge implies that a customer is satisfied with the service; whereas a negative edge implies that they are dissatisfied with or have not used the service. We may be interested in grouping customers and services so that the total and the individual dissatisfaction of customers are minimized.

Definition 1. ***(Local Correlation Clustering)*** *Given an instance of Correlation Clustering $G = (V, E = E^+ \cup E^-)$ and $p \geq 1$, the local objective is to find a partitioning $\mathcal{P}$ that minimizes the ℓ_p norm.*

We use the standard definition of the ℓ_p norm of a vector x: $\|x\|_p = (\sum_u |x_u|^p)^{\frac{1}{p}}$. Since its introduction by Puleo and Milenkovic [40], local objectives for Correlation Clustering have been mainly studied under two models (see Charikar, Gupta, and Schwartz [17], Ahmadi, Khuller, and Saha [1], Kalhan, Makarychev, and Zhou [33]) as in the case of MinDisagree. For the Correlation Clustering on Complete Graphs model, the first approximation algorithm was by Puleo and Milenkovic [40] with an approximation factor of 48 for minimizing the ℓ_p norm. This was later improved to 7 by Charikar et al. [17]. Lastly, Kalhan et al. [33] provided a 5 approximation algorithm. For minimizing the ℓ_∞ norm of the disagreements vector in the Correlation Clustering with Noisy Partial Information model, Charikar et al. [17] provided a $O(\sqrt{n})$ approximation. Kalhan et al. [33] gave an $O(n^{\frac{1}{2}-\frac{1}{2p}} \cdot \log^{\frac{1}{2}+\frac{1}{2p}} n)$-approximation algorithm for minimizing the ℓ_p norm of the disagreements vector. They also provided an almost matching integrality gap of $\Omega(n^{\frac{1}{2}-\frac{1}{2p}})$. For the summary of the results see Table 1.2.

Model	**Complete Unweighted Graphs**	**Arbitrary Weighted Graphs**
Approx. Factor	48 [40] 7 [17] 5 [33]	$O(\sqrt{n})$ (for $p = \infty$) [17] $O\left(n^{\frac{1}{2}-\frac{1}{2p}} \cdot (\log n)^{\frac{1}{2}+\frac{1}{2p}}\right)$ [33]
Integrality Gap	-	$\Omega(n^{\frac{1}{2}-\frac{1}{2p}})$ [33]

Figure 1.2: Known Results for the ℓ_p Objective

1.3 Correlation Clustering with Asymmetric Classification Errors

In this section, we introduce a model for Correlation Clustering that captures both Correlation Clustering on Complete Graphs and Correlation Clustering with Noisy Partial Information [31].

We study the Correlation Clustering problem on complete graphs with edge weights. In our model, the weights on the edges are constrained such that the ratio of the lightest edge in the graph to the heaviest positive edge is at least $\alpha \leq 1$. Thus, if $\boldsymbol{w}$ is the weight of the heaviest positive edge in the graph, then each positive edge has weight in $[\alpha\boldsymbol{w}, \boldsymbol{w}]$ and each negative edge has weight greater than or equal to $\alpha\boldsymbol{w}$. We call this model Correlation Clustering with Asymmetric Classification Errors. Observe that for $\alpha = 1$ our model captures Correlation Clustering on Complete Graphs; as $\alpha \to 0$ its limit is Correlation Clustering with Noisy Partial Information. We argue that Correlation Clustering with Asymmetric Classification Errors is also more adept at capturing the subtleties in real world instances than the two standard models. Indeed, the assumptions made by the Correlation Clustering on Complete Graphs model are too strong, since rarely do real world instances have equal edge weights. In contrast, in the Correlation Clustering with Noisy Partial Information model we can have edge weights that are arbitrarily small or large, an assumption which is too weak. In many real world instances, the edge weights lie in some range $[a, b]$ with $a, b > 0$. Our model captures a larger family of instances.

Definition 2. *Correlation Clustering with Asymmetric Classification Errors is a variant of*

Correlation Clustering on Complete Graphs. We assume that the weight of each positive edge lies in $[\alpha\mathbf{w}, \mathbf{w}]$ *and the weight of each negative edge lies in* $[\alpha\mathbf{w}, \infty)$*, where* $\alpha \in (0, 1]$ *and* $\mathbf{w} > 0$.

Furthermore, the nature of classification errors for objects that are similar and objects that are dissimilar is quite different. In many cases, a *positive* edge uv indicates that the classifier found some actual evidence that u and v are similar; while a negative edge simply means that the classifier could not find any such proof that u and v are similar, it does not mean that the objects u and v are necessarily dissimilar. In some other cases, a *negative* edge uv indicates that the classifier found some evidence that u and v are dissimilar; while a positive edge simply means that the classifier could not find any such proof. We discuss several examples below. Note that in the former case, a positive edge gives a substantially stronger signal than a negative edge and should have a higher weight; in the latter, it is the other way around: a negative edge gives a stronger signal than a positive edge and should have a higher weight. We make this statement more precise in Section 2.1.

The following examples show how the Correlation Clustering with Asymmetric Classification Errors model can help in capturing real world instances. Consider an example from the paper on Correlation Clustering by Pan, Papailiopoulos, Oymak, Recht, Ramchandran, and Jordan [39]. In their experiments, Pan et al. [39] used several data sets including *dblp-2011* and *ENWiki-2013*[1]. In the graph *dblp-2011*, *each vertex represents a scientist and two vertices are connected with an edge if the corresponding authors have co-authored an article.* Thus, a positive edge with weight $\boldsymbol{w}^+$ between Alice and Bob in the Correlation Clustering instance indicates that Alice and Bob are coauthors, which strongly suggests that Alice and Bob work in similar areas of Computer Science. However, it is not true that all researchers working in some area of computer science have co-authored papers with each other. Thus, the negative edge that connects two scientists who do not have an article together does not deserve to

1. These data sets are published by [10, 12, 11, 13]

have the same weight as a positive edge, and thus can be modeled as a negative edge with weight $\boldsymbol{w}^- < \boldsymbol{w}^+$.

Similarly, the vertices of the graph *ENWiki-2013* are Wikipedia pages. Two pages are connected with an edge if there is a link from one page to another. A link from one page to the other is a strong suggestion that the two pages are related and hence can be connected with a positive edge of weight $\boldsymbol{w}^+$, while it is not true that two similar Wikipedia pages necessarily should have a link from one to the other. Thus, it would be better to join such pages with a negative edge of weight $\boldsymbol{w}^- < \boldsymbol{w}^+$.

Consider now the multi-person tracking problem. The problem is modelled as a Correlation Clustering or closely related Lifted Multicut Problem [44, 45] on a graph, whose vertices are people detections in video sequences. Two detections are connected with a positive or negative edge depending on whether the detected people have similar or dissimilar appearance (as well as some other information). In this case, a negative edge (u, v) is more informative since it signals that the classifier has identified body parts that do not match in detections u and v and thus the detected people are likely to be different (a positive edge (u, v) simply indicates that the classifier was not able to find non-matching body parts).

The Correlation Clustering with Asymmetric Classification Errors model captures the examples we discussed above. It is instructive to consider an important special case where all positive edges have weight $\boldsymbol{w}^+$ and all negative edges have weight $\boldsymbol{w}^-$ with $\boldsymbol{w}^+ \neq \boldsymbol{w}^-$. If we were to use the state of the art algorithm for Correlation Clustering on Complete Graphs on our instance for Correlation Clustering with Asymmetric Classification Errors (by completely ignoring edge weights and looking at the instance as an unweighted complete graph), we would get a $\Theta(\max(\boldsymbol{w}^+/\boldsymbol{w}^-, \boldsymbol{w}^-/\boldsymbol{w}^+))$ approximation to the MinDisagree and ℓ_p objectives. While if we were to use the state of the art algorithms for Correlation Clustering with Noisy Partial Information on our instance, we would get a $O(\log n)$ and $\tilde{O}(n^{\frac{1}{2}-\frac{1}{2p}})$ approximations to the MinDisagree and ℓ_p objectives, respectively.

1.4 Our Contribution

1.4.1 Results for the MinDisagree Objective

In Chapter 2, we present an approximation algorithm for the MinDisagree objective in Correlation Clustering with Asymmetric Classification Errors [31]. Our algorithm gives an approximation factor of $A = 3 + 2\log_e 1/\alpha$. Consider the scenario discussed in Section 1.3 where all positive edges have weight $\boldsymbol{w}^+$ and all negative edges have weight $\boldsymbol{w}^-$. If $\boldsymbol{w}^+ \geq \boldsymbol{w}^-$, our algorithm gets a $(3 + 2\log_e \boldsymbol{w}^+/\boldsymbol{w}^-)$ approximation; if $\boldsymbol{w}^+ \leq \boldsymbol{w}^-$, our algorithm gets a 3-approximation.

We note here that the assumption that the weight of positive edges is bounded from above is crucial. Without this assumption (even if we require that negative weights are bounded from above and below), the LP gap is unbounded for every fixed α (this follows from the integrality gap example we present in Theorem 1.4.3). The following is our first main theorem.

Theorem 1.4.1. *There exists a polynomial time $A = 3 + 2\log_e 1/\alpha$ approximation algorithm for Correlation Clustering with Asymmetric Classification Errors.*

We also study a natural extension of our model to the case of complete bipartite graphs. That is, the positive edges across the bipartition have a weight between $[\alpha\boldsymbol{w}, \boldsymbol{w}]$ and the negative edges across the bipartition have a weight of at least $\alpha\boldsymbol{w}$. Note that the state-of-the-art approximation algorithm for Correlation Clustering on Unweighted Complete Bipartite Graphs has an approximation factor of 3 (see Chawla et al. [19]).

Theorem 1.4.2. *There exists a polynomial time $A = 5 + 2\log_e 1/\alpha$ approximation algorithm for Correlation Clustering with Asymmetric Classification Errors on complete bipartite graphs.*

Our next result shows that this approximation ratio is likely best possible for LP-based algorithms. We show this by exhibiting an instance of Correlation Clustering with Asymmetric

Classification Errors such that integrality gap for the natural LP for Correlation Clustering on this instance is $\Omega(\log 1/\alpha)$.

Theorem 1.4.3. *The natural Linear Programming relaxation for Correlation Clustering has an integrality gap of* $\Omega(\log 1/\alpha)$ *for instances of Correlation Clustering with Asymmetric Classification Errors.*

Moreover, we can show that if there is an $o(\log(1/\alpha))$-approximation algorithm whose running time is polynomial in both n and $1/\alpha$, then there is an $o(\log n)-$approximation algorithm for the general weighted case[2](and also for the MultiCut problem). However, we do not know if there is an $o(\log(1/\alpha))-$approximation algorithm for the problem whose running time is polynomial in n and *exponential* in $1/\alpha$. The existence of such an algorithm does not imply that there is an $o(\log n)-$approximation algorithm for the general weighted case (as far as we know).

We show a similar integrality gap result for the Correlation Clustering with Asymmetric Classification Errors on complete bipartite graphs problem.

Theorem 1.4.4. *The natural Linear Programming relaxation for Correlation Clustering has an integrality gap of* $\Omega(\log 1/\alpha)$ *for instances of Correlation Clustering with Asymmetric Classification Errors on complete bipartite graphs.*

1.4.2 Results for the ℓ_p Objective

In Chapter 3, we study the task of minimizing local objectives (Definition 1) under the Correlation Clustering with Asymmetric Classification Errors model (Definition 2). Our main

2. The reduction to the general case works as follows. Consider an instance of Correlation Clustering with arbitrary weights. *Guess* the heaviest edge e that is in disagreement with the optimal clustering. Let $\boldsymbol{w}_e$ be its weight, and set $\boldsymbol{w} = n^2\boldsymbol{w}_e$, and $\alpha = 1/n^4$. Then, assign new weights to all pairs of vertices in the graph. Keep the weights of all edges with weight in the range $[\alpha\boldsymbol{w}, \boldsymbol{w}]$. Set the weights of all edges with weight greater than $\boldsymbol{w}$ to $\boldsymbol{w}$ and the weights of all edges with weight less than $\alpha\boldsymbol{w}$ (including missing edges) to $\alpha\boldsymbol{w}$.

result is an $O\left(\left(\frac{1}{\alpha}\right)^{\frac{1}{2}-\frac{1}{2p}} \cdot \log \frac{1}{\alpha}\right)$ approximation algorithm for minimizing the ℓ_p norm of the disagreements vector, which we now state [32].

Theorem 1.4.5. *There exists a polynomial-time $O\left(\left(\frac{1}{\alpha}\right)^{\frac{1}{2}-\frac{1}{2p}} \cdot \log \frac{1}{\alpha}\right)$-approximation algorithm for the ℓ_p objective in the Correlation Clustering with Asymmetric Classification Errors model.*

For $p = 1$, our algorithm provides an $O(\log \frac{1}{\alpha})$ approximation, which matches the approximation guarantee we presented in Chapter 2 [31] up to constant factors. Consider $p = 2$, that is, the ℓ_2 norm. If we ignored the edge weights and applied the state of the art algorithm in the Correlation Clustering on Complete Graphs model, we would get an $O(\frac{1}{\alpha})$ approximation. If we were to use the state of the art algorithm in the Correlation Clustering with Noisy Partial Information model, we would get an $\tilde{O}\left(n^{1/4}\right)$ approximation. However, by using our algorithm (Theorem 1.4.1), we obtain an $\tilde{O}\left(\left(1/\alpha\right)^{1/4}\right)$ approximation, which is a huge improvement when $1/\alpha \ll n$.

Corollary 1.4.6. *There exists a polynomial-time $O\left(\left(1/\alpha\right)^{1/4} \cdot \log \frac{1}{\alpha}\right)$-approximation algorithm for the ℓ_2 objective in the Correlation Clustering with Asymmetric Classification Errors model.*

Finally, we present the implication of our main result for the ℓ_∞ norm. For the ℓ_∞ norm, Kalhan et al. [33] presented an $\tilde{O}(\sqrt{n})$ approximation under the Correlation Clustering with Noisy Partial Information model. Using our algorithm for Correlation Clustering with Asymmetric Classification Errors we obtain an $\tilde{O}(\sqrt{1/\alpha})$-approximation factor, which is a significant improvement to the approximation guarantee in this setting.

Corollary 1.4.7. *There exists a polynomial-time $O(\sqrt{1/\alpha} \cdot \log 1/\alpha)$-approximation algorithm for the ℓ_∞ objective in the Correlation Clustering with Asymmetric Classification Errors model.*

We emphasize that our approximation ratio for the ℓ_p norm is independent of the graph size and only depends on α.

Our algorithm relies on the natural convex programming relaxation for this problem (Section 3.1). We compliment our positive result (Theorem 1.4.5) by showing that it is likely to be the best possible based on the natural convex program, by providing an almost matching integrality gap.

Theorem 1.4.8. *The natural convex programming relaxation for the ℓ_p objective in the Correlation Clustering with Asymmetric Classification Errors model has an integrality gap of $\Omega\big((1/\alpha)^{\frac{1}{2}-\frac{1}{2p}}\big)$.*

CHAPTER 2

APPROXIMATION ALGORITHM FOR MINDISAGREE

In this chapter, we study MinDisagree in the Correlation Clustering with Asymmetric Classification Errors model. In Section 2.1 we explore the connection between Correlation Clustering with Asymmetric Classification Errors and the semi-random model. In sections 2.2, 2.2.1 we describe our rounding algorithm and the linear programming relaxation. In Section 2.3 we prove our first main result, Theorem 1.4.1. In Section 2.4 we discuss better approximation ratios for values of α appearing in practice. In Section 2.5 we prove our result regarding complete bipartite graphs, Theorem 1.4.2. In Section 2.6 we prove our integrality gap results, Theorem 1.4.3 and Theorem 1.4.4.

2.1 Ground Truth Model

In this section, we formalize the connection between asymmetric classification errors and asymmetric edge weights. For simplicity, we assume that each positive edge has a weight of $\boldsymbol{w}^+$ and each negative edge has a weight of $\boldsymbol{w}^-$. Consider a probabilistic model in which edge labels are assigned by a noisy classifier. Let $\mathcal{C}^* = (C_1^*, \dots C_T^*)$ be the ground truth clustering of the vertex set V. The classifier labels each edge within a cluster with a "+" edge with probability p^+ and as a "−" edge with probability $1 - p^+$; it labels each edge with endpoints in distinct clusters as a "−" edge with probability q^- and as a "+" edge with probability $1 - q^-$. Thus, $(1 - p^+)$ and $(1 - q^-)$ are the classification error probabilities. We assume that all classification errors are independent.

We note that similar models have been previously studied by [7, 26, 37, 5, 35] and others. However, the standard assumption in such models was that the error probabilities, $(1 - p^+)$ and $(1 - q^-)$, are less than a half; that is, $p^+ > 1/2$ and $q^- > 1/2$. Here, we investigate two cases (i) when $p^+ < 1/2 < q^-$ and (ii) when $q^- < 1/2 < p^+$. We assume that $p^+ + q^- > 1$,

which means that the classifier is more likely to connect similar objects with a "+" than dissimilar objects or, equivalently, that the classifier is more likely to connect dissimilar objects with a "−" than similar objects. For instance, consider a classifier that looks for evidence that the objects are similar: if it finds some evidence, it adds a positive edge; otherwise, it adds a negative edge (as described in our examples *dblp-2011* and *ENWiki-2013* in the Introduction). Say, the classifier detects a similarity between two objects in the same ground truth cluster with a probability of only 30% and incorrectly detects similarity between two objects in different ground truth clusters with a probability of 10%. Then, it will add a *negative* edge between two similar objects with probability 70%! While this scenario is not captured by the standard assumption, it is captured by case (i) (here, $p^+ = 0.3 < 1/2 < q^- = 0.9$ and $p^+ + q^- > 1$).

Consider a clustering $\mathcal{C}$ of the vertices. Denote the sets of positive edges and negative edges with both endpoints in the same cluster by $\mathrm{In}^+(\mathcal{C})$ and $\mathrm{In}^-(\mathcal{C})$, respectively, and the sets of positive edges and negative edges with endpoints in different clusters by $\mathrm{Out}^+(\mathcal{C})$ and $\mathrm{Out}^-(\mathcal{C})$, respectively. Then, the log-likelihood function of the clustering $\mathcal{C}$ is,

$$\begin{aligned}
\ell(G;\mathcal{C}) &= \log\Big(\prod_{(u,v)\in \mathrm{In}^+(\mathcal{C})} p^+ \times \prod_{(u,v)\in \mathrm{In}^-(\mathcal{C})}(1-p^+) \times \prod_{(u,v)\in \mathrm{Out}^+(\mathcal{C})}(1-q^-) \times \prod_{(u,v)\in \mathrm{Out}^-(\mathcal{C})} q^-\Big) \\
&= \log\Big((p^+)^{|\mathrm{In}^+(\mathcal{C})|}(1-p^+)^{|\mathrm{In}^-(\mathcal{C})|}\cdot(1-q^-)^{|\mathrm{Out}^+(\mathcal{C})|}(q^-)^{|\mathrm{Out}^-(\mathcal{C})|}\Big) \\
&= |\mathrm{In}^+(\mathcal{C})|\log p^+ + |\mathrm{In}^-(\mathcal{C})|\log(1-p^+) + |\mathrm{Out}^+(\mathcal{C})|\log(1-q^-) + |\mathrm{Out}^-(\mathcal{C})|\log q^- \\
&= \underbrace{\Big(|E^+|\log p^+ + |E^-|\log q^-\Big)}_{\text{constant expression}} - \underbrace{\Big(|\mathrm{Out}^+(\mathcal{C})|\log\frac{p^+}{1-q^-} + |\mathrm{In}^-(\mathcal{C})|\log\frac{q^-}{1-p^+}\Big)}_{\text{MinDisagree objective}}.
\end{aligned}$$

Let $\boldsymbol{w}^+ = \log\frac{p^+}{1-q^-}$ and $\boldsymbol{w}^- = \log\frac{q^-}{1-p^+}$. Then, the negative term – $\Big(|\mathrm{Out}^+(\mathcal{C})|\log\frac{p^+}{1-q^-} + |\mathrm{In}^-(\mathcal{C})|\log\frac{q^-}{1-p^+}\Big)$ – equals $\boldsymbol{w}^+|\mathrm{Out}^+(\mathcal{C})| + \boldsymbol{w}^-|\mathrm{In}^-(\mathcal{C})|$. Note that $|\mathrm{Out}^+(\mathcal{C})|$ is the number of positive edges disagreeing with $\mathcal{C}$ and $|\mathrm{In}^-(\mathcal{C})|$ is the number of negative edges disagreeing with $\mathcal{C}$.

Now observe that the first term in the expression above – $\left(|E^+| \log p^+ + |E^-| \log q^-\right)$ – does not depend on $\mathcal{C}$. It only depends on the instance $G = (V, E^+, E^-)$. Thus, maximizing the log-likelihood function over $\mathcal{C}$ is equivalent to minimizing the following objective

$$\boldsymbol{w}^+(\#\text{ disagreeing ``+''edges}) + \boldsymbol{w}^-(\#\text{ disagreeing ``$-$''edges}).$$

Note that we have $\boldsymbol{w}^+ > \boldsymbol{w}^-$ when $p^+ < 1/2 < q^-$ (case (i) above); in this case, a "+" edge gives a stronger signal than a "−" edge. Similarly, we have $\boldsymbol{w}^- > \boldsymbol{w}^+$ when $q^- < 1/2 < p^+$ (case (ii) above); in this case, a "−" edge gives a stronger signal than a "+" edge.

2.2 Algorithm

In this section, we present an approximation algorithm for Correlation Clustering with Asymmetric Classification Errors. The algorithm first solves a standard LP relaxation and assigns every edge a length of x_{uv} (see Section 2.2.1). Then, one by one it creates new clusters and removes them from the graph. The algorithm creates a cluster C as follows. It picks a random vertex p, called a pivot, among yet unassigned vertices and a random number $R \in [0, 1]$. Then, it adds the pivot p and all vertices u with $f(x_{pu}) \leq R$ to C, where $f : [0, 1] \to [0, 1]$ is a properly chosen function, which we define below. We give a pseudo-code for this algorithm in Algorithm 1.

Our algorithm resembles the LP-based correlation clustering algorithms by Ailon et al. [3] and Chawla et al. [19]. However, a crucial difference between our algorithm and above mentioned algorithms is that our algorithm uses a "dependant" rounding. That is, if for two edges pv_1 and pv_2, we have $f(x_{pv_1}) \leq R$ and $f(x_{pv_2}) \leq R$ at some step t of the algorithm then both v_1 and v_2 are added to the new cluster S_t. The algorithms by Ailon et al. [3] and Chawla et al. [19] make decisions on whether to add v_1 to S_t and v_2 to S_t, independently.

Algorithm 1 Approximation Algorithm

input An instance of Correlation Clustering with Asymmetric Weights $G = (V, E^+, E^-, \boldsymbol{w}_e)$.
  Initialize $t = 0$ and $V_t = V$.
  while $V_t \neq \varnothing$ **do**
    Pick a random pivot $p_t \in V_t$.
    Choose a radius R uniformly at random in $[0, 1]$.
    Create a new cluster S_t; add the pivot p_t to S_t.
    for all $u \in V_t$ **do**
      if $f(x_{p_t u}) \leq R$ **then**
        Add u to S_t.
      end if
    end for
    Let $V_{t+1} = V_t \setminus S_t$ and $t = t + 1$.
  end while
output clustering $\mathcal{S} = (S_0, \ldots, S_{t-1})$.

Also, the choice of the function f is quite different from the functions used by Chawla et al. [19]. In fact, it is influenced by the paper by Garg, Vazirani, and Yannakakis [29].

2.2.1 Linear Programming Relaxation

In this section, we describe a standard linear programming (LP) relaxation for Correlation Clustering which was introduced by Charikar, Guruswami, and Wirth [16]. We first give an integer programming formulation of the Correlation Clustering problem. For every pair of vertices u and v, the integer program (IP) has a variable $x_{uv} \in \{0, 1\}$, which indicates whether u and v belong to the same cluster:

- $x_{uv} = 0$, if u and v belong to the same cluster; and
- $x_{uv} = 1$, otherwise.

We require that $x_{uv} = x_{vu}$, $x_{uu} = 0$ and all x_{uv} satisfy the triangle inequality. That is, $x_{uv} + x_{vw} \geq x_{uw}$.

Every feasible IP solution x defines a partitioning $\mathcal{S} = (S_1, \ldots, S_T)$ in which two vertices u and v belong to the same cluster if and only if $x_{uv} = 0$. A positive edge uv is in disagreement

$$\min \sum_{uv \in E^+} \boldsymbol{w}_{uv} x_{uv} + \sum_{uv \in E^-} \boldsymbol{w}_{uv}(1 - x_{uv}).$$

subject to

$$\begin{aligned} x_{uw} &\leq x_{uv} + x_{vw} && \text{for all } u, v, w \in V \\ x_{uv} &= x_{vu} && \text{for all } u, v \in V \\ x_{uu} &= 0 && \text{for all } u \in V \\ x_{uv} &\in [0, 1] && \text{for all } u, v \in V \end{aligned}$$

Figure 2.1: LP relaxation

with this partitioning if and only if $x_{uv} = 1$; a negative edge uv is in disagreement with this partitioning if and only if $x_{uv} = 0$. Thus, the cost of the partitioning is given by the following linear function:

$$\sum_{uv \in E^+} \boldsymbol{w}_{uv} x_{uv} + \sum_{uv \in E^-} \boldsymbol{w}_{uv}(1 - x_{uv}).$$

We now replace all integrality constraints $x_{uv} \in \{0, 1\}$ in the integer program with linear constraints $x_{uv} \in [0, 1]$. The obtained linear program is given in Figure 2.1. In the paper, we refer to each variable x_{uv} as the length of the edge uv.

2.3 Analysis of the Algorithm

The analysis of our algorithm follows the general approach proposed by Ailon, Charikar, and Newman [3]. Ailon et al. [3] observed that in order to get upper bounds on the approximation factors of their algorithms, it is sufficient to consider how these algorithms behave on triplets of vertices. Below, we present their method adapted to our settings. Then, we will use Theorem 2.3.1 to analyze our algorithm.

Algorithm 2 One iteration of Algorithm 1 on triangle uvw

Pick a random pivot $p \in \{u, v, w\}$.
Choose a random radius R with the uniform distribution in $[0, 1]$.
Create a new cluster S. Insert p in S.
for all $a \in \{u, v, w\} \setminus \{p\}$ **do**
  if $f_\alpha(x_{pa}) \le R$ **then**
    Add a to S .
  end if
end for

2.3.1 General Approach: Triple-Based Analysis

Consider an instance of Correlation Clustering $G = (V, E^+, E^-)$ on three vertices u, v, w. Suppose that the edges uv, vw, and uw have signs $\sigma_{uv}, \sigma_{vw}, \sigma_{uw} \in \{\pm\}$, respectively. We shall call this instance a triangle (u, v, w) and refer to the vector of signs $\sigma = (\sigma_{vw}, \sigma_{uw}, \sigma_{uv})$ as the signature of the triangle (u, v, w).

Let us now assign arbitrary lengths x_{uv}, x_{vw}, and x_{uw} satisfying the triangle inequality to the edges uv, vw, and uw and run one iteration of our algorithm on the triangle uvw (see Algorithm 2).

We say that a positive edge uv is in disagreement with S if $u \in S$ and $v \notin S$ or $u \notin S$ and $v \in S$. Similarly, a negative edge uv is in disagreement with S if $u, v \in S$. Let $cost(u, v \,|\, w)$ be the probability that the edge (u, v) is in disagreement with S given that w is the pivot.

$$cost(u, v \,|\, w) = \begin{cases} \Pr(u \in S, v \notin S \text{ or } u \notin S, v \in S \,|\, p = w), & \text{if } \sigma_{uv} = \text{“+”}; \\ \Pr(u \in S, v \in S \,|\, p = w), & \text{if } \sigma_{uv} = \text{“}-\text{”}. \end{cases}$$

Let $lp(u, v \,|\, w)$ be the LP contribution of the edge (u, v) times the probability of it being removed, conditioned on w being the pivot.

$$lp(u, v \,|\, w) = \begin{cases} x_{uv} \cdot \Pr(u \in S \text{ or } v \in S \,|\, p = w), & \text{if } \sigma_{uv} = \text{“+”}; \\ (1 - x_{uv}) \cdot \Pr(u \in S \text{ or } v \in S \,|\, p = w), & \text{if } \sigma_{uv} = \text{“}-\text{”}. \end{cases}$$

We now define two functions $ALG^{\sigma}(x, y, z)$ and $LP^{\sigma}(x, y, z)$. To this end, construct a triangle (u, v, w) with signature σ edge lengths x, y, z (where $x_{vw} = x$, $x_{uw} = y$, $x_{uv} = z$). Then,

$$ALG^{\sigma}(x, y, z) = \boldsymbol{w}_{uv} \cdot cost(u, v \,|\, w) + \boldsymbol{w}_{uw} \cdot cost(u, w \,|\, v) + \boldsymbol{w}_{vw} \cdot cost(v, w \,|\, u);$$
$$LP^{\sigma}(x, y, z) = \boldsymbol{w}_{uv} \cdot lp(u, v \,|\, w) + \boldsymbol{w}_{uw} \cdot lp(u, w \,|\, v) + \boldsymbol{w}_{vw} \cdot lp(v, w \,|\, u).$$

We will use the following theorem from the paper by Chawla, Makarychev, Schramm, and Yaroslavtsev [19] (Lemma 4) to analyze our algorithm. This theorem was first proved by Ailon, Charikar, and Newman [3] but it was not stated in this form in their paper.

Theorem 2.3.1 (see [3] and [19])**.** *Consider a function f_{α} with $f_{\alpha}(0) = 0$. If for all signatures $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ (where each $\sigma_i \in \{\pm\}$) and edge lengths x, y, and z satisfying the triangle inequality, we have $ALG^{\sigma}(x, y, z) \leq \rho LP^{\sigma}(x, y, z)$, then the approximation factor of the algorithm is at most ρ.*

2.3.2 Analysis of the Approximation Algorithm

Proof of Theorem 1.4.1. Without loss of generality we assume that the scaling parameter $\boldsymbol{w}$ is 1. We use different functions for $\alpha \leq 0.169$ and $\alpha \geq 0.169$. Let $A = 3 + 2 \log_e 1/\alpha$. For $\alpha \leq 0.169$, we define $f_{\alpha}(x)$ as follows (see Figure 2.2):

$$f_{\alpha}(x) = \begin{cases} 1 - e^{-Ax}, & \text{if } 0 \leq x < \frac{1}{2} - \frac{1}{2A}; \\ 1, & \text{otherwise;} \end{cases}$$

and, for $\alpha \geq 0.169$, we define $f_\alpha(x)$ as follows:

$$f_\alpha(x) = \begin{cases} 0, & \text{if } x < \frac{1}{A} \\ \frac{1-\alpha}{3}, & \text{if } \frac{1}{A} \leq x < \frac{1}{2} - \frac{1}{2A} \\ 1, & \text{if } x \geq \frac{1}{2} - \frac{1}{2A} \end{cases}$$

Our analysis of the algorithm relies on Theorem 2.3.1. We will show that for every triangle (u_1, u_2, u_3) with edge lengths (x_1, x_2, x_3) (satisfying the triangle inequality) and signature $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, we have

$$ALG^\sigma(x_1, x_2, x_3) \leq A \cdot LP^\sigma(x_1, x_2, x_3). \tag{2.1}$$

Therefore, by Theorem 2.3.1, our algorithm gives an A-approximation.

Without loss of generality, we assume that $x_1 \leq x_2 \leq x_3$. When $i \in \{1, 2, 3\}$ is fixed, we will denote the other two elements of $\{1, 2, 3\}$ by k and j, so that $j < k$. For $i \in \{1, 2, 3\}$, let $e_i = (u_j, u_k)$ (the edge opposite to u_i), $w_i = \boldsymbol{w}_{e_i}$, $x_i = x_{u_j u_k}$, $y_i = f_\alpha(x_i)$, and

$$t_i = A \cdot lp(u_j, u_k | u_i) - cost(u_j, u_k | u_i).$$

Observe that (2.1) is equivalent to the inequality $w_1 t_1 + w_2 t_2 + w_3 t_3 \geq 0$. We now prove that this inequality always holds.

Lemma 2.3.2. *We have*

$$w_1 t_1 + w_2 t_2 + w_3 t_3 \geq 0 \tag{2.2}$$

We express each t_i in terms of x_i's and y_i's.

Claim 2.3.3. *For every $i \in \{1,2,3\}$, we have*

$$t_i = \begin{cases} A(1-y_j)x_i - (y_k - y_j), & \text{if } \sigma_i = \text{“+”} \\ A(1-y_j)(1-x_i) - (1-y_k), & \text{if } \sigma_i = \text{“−”} \end{cases}$$

Proof. If $\sigma_i =$ “+”, then

$$\begin{aligned} t_i &= A \cdot lp(u_j, u_k | u_i) - cost(u_j, u_k | u_i) \\ &= Ax_{u_j u_k} \cdot \Pr(u_j \in S \text{ or } u_k \in S \,|\, p = u_i] - \Pr(u_j \in S, u_k \notin S \text{ or } u_j \notin S, u_k \in S \,|\, p = u_i] \\ &= Ax_i \cdot \Pr(f_\alpha(x_k) \le R \text{ or } f_\alpha(x_j) \le R) - \Pr(f_\alpha(x_k) \le R < f_\alpha(x_j) \text{ or } f_\alpha(x_j) \le R < f_\alpha(x_k)) \\ &= Ax_i(1-y_j) - (y_k - y_j), \end{aligned}$$

where we used that $y_k = f_\alpha(x_k) \ge f_\alpha(x_j) = y_j$ (since $x_k \ge x_j$ and $f_\alpha(x)$ is non-decreasing).

If $\sigma_i =$ “−”, then similarly to the previous case, we have

$$\begin{aligned} t_i &= A \cdot lp(u_j, u_k | u_i) - cost(u_j, u_k | u_i) \\ &= A(1 - x_{u_j u_k}) \cdot \Pr(u_j \in S \text{ or } u_k \in S \,|\, p = u_i) - \Pr(u_j \in S, u_k \in S \,|\, p = w) \\ &= A(1-x_i) \cdot \Pr(f_\alpha(x_k) \le R \text{ or } f_\alpha(x_j) \le R) - \Pr(f_\alpha(x_k) \le R, f_\alpha(x_j) \le R) \\ &= A(1-x_i) \cdot (1-y_j) - (1-y_k). \end{aligned}$$

□

We say that edge e_i *pays for itself* if $t_i \ge 0$. Note that if all edges e_1, e_2, e_3 pay for themselves then the desired inequality (2.2) holds. First, we show that all negative edges pay for themselves.

Claim 2.3.4. *If $\sigma_i =$ “−”, then $t_i \ge 0$.*

Proof. By Claim 2.3.3, $t_i = A(1-y_j)(1-x_i) - 1 - y_k$. Thus, we need to show that $A(1-y_j)(1-x_i) \geq 1 - y_k$. If $x_k \geq \frac{1}{2} - \frac{1}{2A}$ then $y_k = 1$, and the inequality trivially holds. If $x_k < \frac{1}{2} - \frac{1}{2A}$, then using $x_j \leq x_k$, we get

$$A > \frac{1}{1-2x_k} \geq \frac{1}{1-x_k-x_j} \geq \frac{1}{1-x_i},$$

here we used the triangle inequality $x_k + x_j \geq x_i$. Thus

$$A(1-y_j)(1-x_i) \geq A(1-y_k)(1-x_i) \geq 1 - y_k.$$

□

We now show that for short edges e_i, it is sufficient to consider only the case when $\sigma_i =$ "+". Specifically, we prove the following claim.

Claim 2.3.5. *Suppose that $x_i < \frac{1}{2} - \frac{1}{2A}$. If (2.2) holds for σ with $\sigma_i =$ "+", then (2.2) also holds for σ' obtained from σ by changing the sign of σ_i to "−".*

Proof. To prove the claim, we show that the value of t_i is greater for σ' than for σ. That is,

$$A(1-y_j)x_i - (y_k - y_j) < A(1-y_j)(1-x_i) - (1-y_k).$$

Note that the values of t_j and t_k do not depend on σ_i and thus do not change if we replace σ with σ'. Since f_α is non-decreasing and $x_j \leq x_k$, we have $y_j \leq y_k$. Hence,

$$x_i < \frac{1}{2} - \frac{1}{2A} = \frac{1}{2} + \frac{1}{2A} - \frac{1}{A} \leq \frac{1}{2} + \frac{1}{2A} - \frac{(1-y_k)}{A(1-y_j)}.$$

Thus,

$$2A(1-y_j)x_i < A(1-y_j) + 1 - y_j - 2(1-y_k).$$

Therefore,

$$A(1-y_j)x_i - (y_k - y_j) < A(1-y_j)(1-x_i) - (1-y_k),$$

as required. □

Unlike negative edges, positive edges do not necessarily pay for themselves. We now prove that positive edges of length at least $1/A$ pay for themselves.

Claim 2.3.6. *If $\sigma_i =$ "+" and $x_i \geq 1/A$, then $t_i \geq 0$.*

Proof. We have,

$$t_i = A(1-y_j)x_i - (y_k - y_j) \geq (1-y_j) - (y_k - y_j) = 1 - y_k \geq 0.$$

□

We now separately consider two cases $\alpha \leq 0.169$ and $\alpha \geq 0.169$.

2.3.3 Analysis of the Approximation Algorithm for $\alpha \leq 0.169$

First, we consider the case of $\alpha \leq 0.169$.

Proof of Lemma 2.3.2 for $\alpha \leq 0.169$. We first show that if $x_3 < \frac{1}{2} - \frac{1}{2A}$, then all three edges e_1, e_2, and e_3 pay for themselves.

Claim 2.3.7. *If $x_3 < \frac{1}{2} - \frac{1}{2A}$, then $t_i \geq 0$ for every i.*

Proof. Since $x_3 < \frac{1}{2} - \frac{1}{2A}$, for every $i \in \{1,2,3\}$ we have $x_i < \frac{1}{2} - \frac{1}{2A}$ and thus $y_i \equiv f_\alpha(x_i) = 1 - e^{-Ax_i}$. We show that $t_i \geq 0$ for all i. Fix i. If $\sigma_i =$ "$-$", then, by Claim 2.3.4, $t_i \geq 0$. If $\sigma_i =$ "+", then

$$y_k - y_j = e^{-Ax_j} - e^{-Ax_k} = e^{-Ax_j}\left(1 - e^{-A(x_k - x_j)}\right) \leq$$
$$\leq e^{-Ax_j} A(x_k - x_j) \leq e^{-Ax_j} Ax_i = A(1-y_j)x_i,$$

where the first inequality follows from the inequality $1 - e^{-x} \leq x$, and the second inequality follows from the triangle inequality. Thus, $t_i = A(1 - y_j)x_i - (y_k - y_j) \geq 0$. $\square$

We conclude that if $x_3 < \frac{1}{2} - \frac{1}{2A}$, then (2.2) holds. The case $x_3 < \frac{1}{2} - \frac{1}{2A}$ is the most interesting case in the analysis; the rest of the proof is more technical. As a side note, let us point out that Theorem 1.4.1 has dependence $A = 3 + 2\log_e 1/\alpha$ because (i) $f_\alpha(x)$ must be equal to $C - e^{-Ax}$ or a slower growing function so that Claim 2.3.7 holds (ii) Theorem 2.3.1 requires that $f_\alpha(0) = 0$, and finally (iii) we will need below that $1 - f\left(\frac{1}{2} - \frac{3}{2A}\right) \leq \alpha$.

From now on, we assume that $x_3 \geq \frac{1}{2} - \frac{1}{2A}$ and, consequently, $y_3 = f(x_3) = 1$. Observe that if $x_1 \geq \frac{1}{A}$, then all $x_i \geq \frac{1}{A}$ and thus, by Claims 2.3.4 and 2.3.6, all $t_i \geq 0$ and we are done. Similarly, if $x_2 \geq \frac{1}{2} - \frac{1}{2A}$, then $x_2 \geq \frac{1}{A}$ (since $A \geq 3$). Hence, $t_2 \geq 0$ and $t_3 \geq 0$; additionally, $y_2 = y_3 = 1$. Thus $t_1 = 0$ and inequality (2.2) holds. Therefore, it remains to show that inequality (2.2) holds when

$$x_1 < \frac{1}{A}, \quad x_2 < \frac{1}{2} - \frac{1}{2A}, \text{ and } x_3 \geq \frac{1}{2} - \frac{1}{2A}.$$

By Claim 2.3.5, we may also assume that $\sigma_1 =$ "+" and $\sigma_2 =$ "+". Since $\alpha \leq 0.169$, we have $A > 5$ and

$$x_2 \geq x_3 - x_1 \geq \left(\frac{1}{2} - \frac{1}{2A}\right) - \frac{1}{A} > \frac{1}{A} \text{ and } x_3 \geq \frac{1}{2} - \frac{1}{2A} > \frac{1}{A}.$$

Thus, by Claims 2.3.4 and 2.3.6, $t_2 \geq 0$ and $t_3 \geq 0$. Hence, $w_2 t_2 + w_3 t_3 \geq \alpha(w_2 + w_3)$. Also, recall that e_1 is a positive edge and thus $w_1 \leq 1$. Therefore, it is sufficient to show that

$$t_1 \geq -\alpha(t_2 + t_3). \tag{2.3}$$

Now we separately consider two possible signatures $\sigma =$ ("+", "+", "+") and $\sigma =$ ("+", "+", "−").

First, assume that $\sigma = (\text{“+”}, \text{“+”}, \text{“+”})$. We need to show that

$$A(1-y_2)x_1 - (1-y_2) \geq \alpha\Big((1-y_1) + (y_2-y_1) - A(1-y_1)x_2 - A(1-y_1)x_3\Big).$$

Here, we used that $y_3 = 1$. Note that $x_2 \geq x_3 - x_1 \geq \frac{1}{2} - \frac{1}{2A} - \frac{1}{A} = \frac{1}{2} - \frac{3}{2A}$. Therefore,

$$1 - y_2 \leq 1 - \Big(1 - e^{-A(\frac{1}{2} - \frac{3}{2A})}\Big) = e^{-\frac{3}{2} - \log_e \frac{1}{\alpha} + \frac{3}{2}} = e^{-\log_e \frac{1}{\alpha}} = \alpha.$$

Thus, $(1-y_2) + \alpha(1-y_1) + \alpha(y_2 - y_1) \leq \alpha y_2 + 2\alpha(1-y_1)$. To finish the analysis of the case $\sigma = (\text{“+”}, \text{“+”}, \text{“+”})$, it is sufficient to show that

$$\alpha y_2 + 2\alpha(1-y_1) \leq A(1-y_2)x_1 + \alpha A(1-y_1)x_2 + \alpha A(1-y_1)x_3.$$

This inequality immediately follows from the following claim (we simply need to add up (2.4) and (2.5) and multiply the result by α).

Claim 2.3.8. *For $c = 0.224$, we have*

$$(2-c)(1-y_1) \leq A(1-y_1)x_2; \text{ and} \tag{2.4}$$

$$y_2 + c(1-y_1) \leq A(1-y_1)x_3. \tag{2.5}$$

Proof. Since $c \geq 2 - \log_e \frac{1}{0.169} \geq 2 - \log_e \frac{1}{\alpha}$ (recall that $\alpha \leq 0.169$), we have

$$2 - c \leq \log_e \frac{1}{\alpha} = \frac{A}{2} - \frac{3}{2} \leq Ax_2.$$

Therefore, (2.4) holds. We also have,

$$c \leq 0.169 + \log_e \frac{1}{0.169} + 1 - e \leq \alpha + \log_e \frac{1}{\alpha} + 1 - e.$$

Thus, $e-\alpha \leq \frac{A}{2}-\frac{1}{2}-c \leq Ax_3-c$. Therefore,

$$e^{-1}\left(Ax_3-c\right) \geq 1-\alpha e^{-1} = 1-e^{-A\left(\frac{1}{2}-\frac{1}{2A}\right)} \geq y_2, \tag{2.6}$$

where we used that $x_2 < \frac{1}{2}-\frac{1}{2A}$ and $y_2 = f_\alpha(x_2) = 1-e^{-Ax_2}$. Observe that from inequalities (2.6) and $x_1 < \frac{1}{A}$ it follows that

$$y_2 \leq \left(1-f\Big(\frac{1}{A}\Big)\right)\left(Ax_3-c\right) \leq (1-y_1)(Ax_3-c),$$

which implies (2.5). □

Now, assume that $\sigma = (\text{“+”}, \text{“+”}, \text{“}-\text{”})$. We need to prove the following inequality,

$$(1-y_2)+\alpha(1-y_1+1-y_2) \leq A(1-y_2)x_1+\alpha A(1-y_1)(x_2+1-x_3). \tag{2.7}$$

As before,

$$(1-y_2)+\alpha(1-y_1+1-y_2) \leq \alpha+\alpha(1-y_1+1-y_2) \leq \alpha+2\alpha(1-y_1). \tag{2.8}$$

On the other hand,

$$\begin{aligned} A(1-y_2)x_1+\alpha A(1-y_1)(x_2+1-x_3) &\geq \alpha A(1-y_1)(1-x_1+x_1+x_2-x_3) \\ &\geq \alpha A(1-y_1)(1-x_1) \\ &\geq \alpha A(1-y_1)\left(1-\frac{1}{A}\right) \\ &= \alpha(1-y_1)(A-1) \end{aligned} \tag{2.9}$$

where the second inequality is due to the triangle inequality, and the third inequality is due

to $x_1 < \frac{1}{A}$. Finally, observe that $1 \leq 2e^{-1} \log_e \frac{1}{\alpha} = e^{-1}(A-3) \leq (1-y_1)(A-3)$. We get,

$$\alpha(1-y_1)(A-1) \geq \alpha + 2\alpha(1-y_1). \tag{2.10}$$

Combining (2.8), (2.9), and (2.10), we get (2.7). This concludes the case analysis and the proof of Theorem 1.4.1 for the regime $\alpha \leq 0.169$.

2.3.4 Analysis of the Approximation Algorithm for $\alpha \geq 0.169$

We now consider the case when $\alpha \geq 0.169$. Observe that for $\alpha \geq 0.169$

$$A = 3 + 2\log_e(1/\alpha) \geq \frac{6\alpha + 3 - (1-\alpha)^2}{3\alpha} \tag{2.11}$$

and

$$\frac{1-\alpha}{3} \leq \frac{2\alpha}{1+\alpha} \tag{2.12}$$

Proof of Lemma 2.3.2 for $\alpha \geq 0.169$. Observe that if $x_1 \geq \frac{1}{A}$, then all $x_i \geq 1/A$ and thus, by Claims 2.3.4 and 2.3.6, all $t_i \geq 0$ and we are done. Moreover, if $x_3 < \frac{1}{A}$ then all $x_i < 1/A$ implying $y_i = 0$ and thus, $t_i \geq 0$ for $\sigma_i =$ " $+$ ". This combined with Claim 2.3.4 imply all $t_i \geq 0$ and we are done. Similarly, if $x_2 \geq \frac{1}{2} - \frac{1}{2A}$, then $x_2 \geq 1/A$ (since $A \geq 3$). Hence, $t_2 \geq 0$ and $t_3 \geq 0$; additionally, we have $y_2 = y_3 = 1$. Thus, $t_1 = 0$ and we are done.

Therefore, we will assume below that

$$x_1 < \frac{1}{A}, \;\; x_2 < \frac{1}{2} - \frac{1}{2A}, \;\; x_3 \geq \frac{1}{A}.$$

Furthermore, by Claim 2.3.5, we may assume $\sigma_1 =$ "+" and $\sigma_2 =$ "+". We consider four cases: (i) $x_2 \geq 1/A,\ x_3 \geq 1/2 - 1/(2A)$, (ii) $x_2 < 1/A,\ x_3 \geq 1/2 - 1/(2A)$, (iii) $x_2 \geq 1/A,\ x_3 < 1/2 - 1/(2A)$, and (iv) $x_2 < 1/A,\ x_3 < 1/2 - 1/(2A)$.

Consider the case $x_2 \geq \frac{1}{A},\ x_3 \geq \frac{1}{2} - \frac{1}{2A}$. Then $y_1 = 0,\ y_2 = (1-\alpha)/3,\ y_3 = 1$. By

Claims 2.3.4 and 2.3.6, $t_2, t_3 \geq 0$, and e_2, e_3 pay for themselves. If $t_1 \geq 0$, we are done. So we will assume below that $t_1 < 0$. Then,

$$w_1 t_1 + w_2 t_2 + w_3 t_3 \geq 1 \cdot t_1 + \alpha t_2 + \alpha t_3 \tag{2.13}$$

(recall that we assume that e_1 is a positive edge and thus $w_1 \leq 1$).

Now we separately consider two possible signatures $\sigma = (\text{“+”}, \text{“+”}, \text{“+”})$ and $\sigma = (\text{“+”}, \text{“+”}, \text{“−”})$.

First, assume that $\sigma = (\text{“+”}, \text{“+”}, \text{“+”})$. Because of (2.13), to prove (2.2) it is sufficient to show

$$(1 - y_2) + \alpha + \alpha y_2 \leq A(1 - y_2)x_1 + \alpha A x_2 + \alpha A x_3 \tag{2.14}$$

From (2.11) it follows that

$$1 + \alpha \leq \frac{(1-\alpha)^2}{3} + \alpha(A - 1)$$

which implies (2.15) due to $x_3 \geq \frac{1}{2} - \frac{1}{2A}$

$$1 + \alpha \leq \frac{(1-\alpha)^2}{3} + 2\alpha A x_3 \tag{2.15}$$

Observe that from (2.15) together with triangle inequality and $y_2 = \frac{1-\alpha}{3} \leq 1 - \alpha$ it follows that

$$1 + \alpha \leq (1 - \alpha)y_2 + A(1 - y_2)x_1 - \alpha A x_1 + \alpha A x_1 + \alpha A x_2 + \alpha A x_3$$

which is equivalent to (2.14).

Now, assume that $\sigma = (\text{“+”}, \text{“+”}, \text{“−”})$. Because of (2.13), to prove (2.2) it is sufficient

to show

$$(1-y_2)+\alpha+\alpha(1-y_2) \le A(1-y_2)x_1 + \alpha A x_2 + \alpha A(1-x_3) \tag{2.16}$$

From (2.11) and $y_2 = \frac{1-\alpha}{3}$ it follows that

$$1+2\alpha \le \frac{(1-\alpha)^2}{3} + \alpha A \le y_2(1+\alpha) + \alpha A$$

Since $y_2 \le 1-\alpha$,

$$(1+2\alpha) \le (1+\alpha)y_2 + A(1-y_2)x_1 - \alpha A x_1 + \alpha A,$$

Hence, using the triangle inequality,

$$1+2\alpha \le (1+\alpha)y_2 + A(1-y_2)x_1 - \alpha A x_1 + \alpha A + \alpha A x_1 + \alpha A x_2 - \alpha A x_3.$$

which is equivalent to (2.16).

Consider the case $x_2 < \frac{1}{A},\ x_3 \ge \frac{1}{2} - \frac{1}{2A}$. Then $y_1 = y_2 = 0,\ y_3 = 1$. Observe that $t_3 \ge 0$ and $t_1, t_2 < 0$. Then,

$$w_1 t_1 + w_2 t_2 + w_3 t_3 \ge 1 \cdot t_1 + 1 \cdot t_2 + \alpha t_3. \tag{2.17}$$

(recall that we assume that e_1, e_2 are positive edges and thus $w_1, w_2 \le 1$). Furthermore, since $x_3 \ge \frac{1}{2} - \frac{1}{2A}$ we have

$$Ax_3 \ge A(1-x_3) - 1. \tag{2.18}$$

From (2.18), we get that if (2.2) holds for σ with $\sigma_3 =$ "$-$", then (2.2) also holds for σ' obtained from σ by changing the sign of σ_3 to "+". Thus without loss of generality $\sigma_3 =$ "$-$" and we only need to consider $\sigma =$ ("+", "+", "$-$"). Then, because of (2.17), to prove (2.2) it

is sufficient to show

$$1 + 1 + \alpha \leq Ax_1 + Ax_2 + \alpha A(1 - x_3). \tag{2.19}$$

From (2.11) it follows that

$$A \geq \frac{5 + \alpha}{\alpha + 1}$$

which is equivalent to

$$2 + \alpha \leq \alpha A + (1 - \alpha)(\frac{A}{2} - \frac{1}{2}). \tag{2.20}$$

Observe that from (2.20) together with triangle inequality and $x_3 \geq \frac{1}{2} - \frac{1}{2A}$ it follows that

$$2 + \alpha \leq \alpha A + (1 - \alpha)Ax_3 = Ax_3 + \alpha A(1 - x_3) \leq Ax_1 + Ax_2 + \alpha A(1 - x_3).$$

Consider the case $x_2 \geq \frac{1}{A},\ x_3 < \frac{1}{2} - \frac{1}{2A}$**.** Then $y_1 = 0,\ y_3 = {}^{(1-\alpha)}/_{3}$. By Claim 2.3.5 we only need to consider $\sigma =$ ("+", "+", "+"). Then by Claim 2.3.6, $t_2, t_3 \geq 0$. Thus, if $t_1 \geq 0$ then $w_1t_1 + w_2t_2 + w_3t_3 \geq 0$. Let us assume that $t_1 < 0$. Since e_1 is a positive edge, we have $w_1 \leq 1$. Thus,

$$w_1t_1 + w_2t_2 + w_3t_3 \geq 1 \cdot t_1 + \alpha t_2 + \alpha t_3$$

Now to prove (2.2) it is sufficient to show

$$y_3 - y_2 + \alpha y_3 + \alpha y_2 \leq A(1 - y_2)x_1 + \alpha Ax_2 + \alpha Ax_3 \tag{2.21}$$

Observe that since $x_3 \geq \frac{1}{A}$ we have

$$2\alpha \leq (1 - \alpha)y_2 + 2\alpha Ax_3. \tag{2.22}$$

Inequalities (2.22) and (2.12) imply

$$(1 + \alpha)y_3 \leq (1 - \alpha)y_2 + 2\alpha Ax_3. \tag{2.23}$$

Observe that from (2.23) together with triangle inequality and $y_2 \leq 1-\alpha$ it follows that

$$(1+\alpha)y_3 \leq (1-\alpha)y_2 + A(1-y_2)x_1 - \alpha A x_1 + \alpha A x_1 + \alpha A x_2 + \alpha A x_3$$

which is equivalent to (2.21).

Consider the case $x_2 < \frac{1}{A},\ x_3 < \frac{1}{2} - \frac{1}{2A}$. Then $y_1 = y_2 = 0$. By Claim 2.3.5 we only need to consider $\sigma = (\text{“+”}, \text{“+”}, \text{“+”})$. Then by Claim 2.3.6, $t_3 \geq 0$.

If $x_1 \geq y_3/A$ then $t_1, t_2 \geq 0$ and we are done. Thus we assume $x_1 < y_3/A$ which implies $t_1 < 0$. We consider two different regimes: (i) $x_2 \geq y_3/A$ and (ii) $x_2 < y_3/A$.

First, assume that $x_2 \geq y_3/A$ which implies $t_2 \geq 0$. Then,

$$w_1 t_1 + w_2 t_2 + w_3 t_3 \geq 1 \cdot t_1 + \alpha t_2 + \alpha t_3 \tag{2.24}$$

(recall that we assume that e_1 is a positive edge and thus $w_1 \leq 1$).

Because of (2.24), to prove (2.2) it is sufficient to show

$$y_3 + \alpha y_3 \leq A x_1 + \alpha A x_2 + \alpha A x_3 \tag{2.25}$$

Observe that by (2.12) and $y_3 = (1-\alpha)/3$ we have

$$(1+\alpha)y_3 \leq 2\alpha \leq 2\alpha A x_3 \leq \alpha A x_3 + \alpha A x_1 + \alpha A x_2 \leq A x_1 + \alpha A x_2 + \alpha A x_3$$

where the second inequality follows from $x_3 \geq \frac{1}{A}$ and the third inequality follows from triangle inequality.

Now, assume that $x_2 < y_3/A$ which implies $t_2 < 0$. Then,

$$w_1 t_1 + w_2 t_2 + w_3 t_3 \geq 1 \cdot t_1 + 1 \cdot t_2 + \alpha t_3 \tag{2.26}$$

(recall that we assume that e_1, e_2 are positive edges and thus $w_1, w_2 \leq 1$).

Because of (2.26), to prove (2.2) it is sufficient to show

$$2y_3 \leq Ax_1 + Ax_2 + \alpha Ax_3 \tag{2.27}$$

Observe that by (2.12) and $x_3 \geq \frac{1}{A}$

$$2y_3 \leq \frac{4\alpha}{1+\alpha} \leq 1 + \alpha \leq (1+\alpha)Ax_3 \leq Ax_1 + Ax_2 + \alpha Ax_3$$

where the last inequality follows from triangle inequality.

This concludes the case analysis and the proof of Theorem 1.4.1 for the regime $\alpha \geq 0.169$. □

2.4 Better approximation for values of α appearing in practice

We note that the choice of function $f(x)$ in Theorem 1.4.1 is somewhat suboptimal. However, for every $\alpha \in (0, 1]$, we can compute the optimal function $f_{opt}(x)$ (with high precision) using linear programming. Using this function f_{opt}, we can achieve an approximation factor A_{opt} better than the approximation factor $A_{thm} = 3 + 2\log_e 1/\alpha$ guaranteed by Theorem 1.4.1.[1] While asymptotically $A_{thm}/A_{opt} \to 1$ as $\alpha \to 0$, A_{opt} is noticeably better than A_{thm} for many values of α that are likely to appear in practice (say, for $\alpha \in (10^{-8}, 0.1)$). We list approximation factors A_{thm} and A_{opt} for several values of α in Table 2.1; we also plot the dependence of A_{thm} and A_{opt} on α in Figure 2.3.

1. It is also possible to slightly modify Algorithm 1 so that it gets approximation A_{opt} without explicitly computing f. We omit the details here.

Table 2.1: Approximation factors A_{thm} and A_{opt} for different α-s.

$\log_e 1/\alpha$	$1/\alpha$	A_{thm}	A_{opt}
0	1	3	3
1.61	5	6.22	4.32
2.30	10	7.61	4.63
3.91	50	10.82	6.07
4.61	100	12.21	6.78
6.21	500	15.43	8.69
6.91	1000	16.82	9.62
8.52	5 000	20.03	11.9
10	22 026.5	23	14.2
15	3.3×10^6	33	22.6
20	4.9×10^8	43	31.3

2.5 Analysis of the Algorithm for Complete Bipartite Graphs

Proof of Theorem 1.4.2. The proof is similar to the proof of Theorem 1.4.1. Without loss of generality we assume that the scaling parameter $\boldsymbol{w}$ is 1. Define $f(x)$ as follows

$$f(x) = \begin{cases} 1 - e^{-Ax}, & \text{if } 0 \leq x < \frac{1}{2} - \frac{1}{2A} \\ 1, & \text{otherwise} \end{cases}$$

where $A = 5 + 2\log_e 1/\alpha$. Our analysis of the algorithm relies on Theorem 2.3.1. Since in the proof of Theorem 2.3.1, we assumed that all edges are present, let us add missing edges (edges inside parts) to the bipartite graph and assign them weight 0; to be specific, we assume that they are positive edges. (It is important to note that Theorem 2.3.1 is true even when edges have zero weights). We will still refer to these edges as 'missing edges'.

We will show that for every triangle (u_1, u_2, u_3) with edge lengths (x_1, x_2, x_3) (satisfying the triangle inequality) and signature $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, we have

$$ALG^{\sigma}(x_1, x_2, x_3) \leq A \cdot LP^{\sigma}(x_1, x_2, x_3) \tag{2.28}$$

Therefore, by Theorem 2.3.1, our algorithm gives an A-approximation. In addition to Theorem 2.3.1 we use Claims 2.3.3, 2.3.4, 2.3.5, 2.3.6 and 2.3.7. Recall that proofs of these claims rely on f being non-decreasing which is satisfied by the above choice. Observe that (2.28) is equivalent to

$$\sum_{i=1}^{3} w_i t_i \geq 0. \tag{2.29}$$

Observe that if $x_1 \geq \frac{1}{A}$, then all $x_i \geq \frac{1}{A}$ and thus, by Claims 2.3.4 and 2.3.6, all $t_i \geq 0$ and we are done. Similarly, if $x_2 \geq \frac{1}{2} - \frac{1}{2A} \geq \frac{1}{A}$ (since $A > 3$), then $t_2 \geq 0$ and $t_3 \geq 0$; additionally, $y_2 = y_3 = 1$, thus $t_1 = 0$ and we are done. Furthermore, if $x_3 < \frac{1}{2} - \frac{1}{2A}$ then all $x_i < \frac{1}{2} - \frac{1}{2A}$ and thus, by Claim 2.3.7, all $t_i \geq 0$ and we are done. Therefore, we will assume below that $x_1 < \frac{1}{A}$, $x_2 < \frac{1}{2} - \frac{1}{2A}$, and $x_3 \geq \frac{1}{2} - \frac{1}{2A}$. Further, by the triangle inequality $x_2 \geq x_3 - x_1 \geq \frac{A-1}{2A} - x_1 \geq \frac{A-3}{2A}$. We have (here we use that $A \geq 5$),

$$x_1 \leq \frac{1}{A} \leq \frac{A-3}{2A} \leq \frac{A-1}{2A} - x_1 \leq x_2 < \frac{A-1}{2A} \leq x_3 \leq x_1 + x_2.$$

We will use below that

$$e^{A(x_2-x_1)} \geq e^{A(\frac{A-1}{2A}-2x_1)} = e^{2+\log\frac{1}{\alpha}-2Ax_1} = e^{2(1-Ax_1)}/\alpha \geq 1/\alpha.$$

By Claim 2.3.5, we may also assume that $\sigma_1 =$ "+" and $\sigma_2 =$ "+" (and since we assume that missing edges are positive). By Claims 2.3.4 and 2.3.6, $t_2 \geq 0$ and $t_3 \geq 0$ (edges e_2 and e_3 pay for themselves). If $t_1 \geq 0$, we are done. So we will assume below that $t_1 < 0$. Since G is a complete bipartite graph, a triangle (u_1, u_2, u_3) contains either (i) no edges or (ii) two edges. In case (i) we have $w_1 = w_2 = w_3 = 0$ and (2.29) holds trivially. In case (ii) if e_1 is the missing edge then $w_1 = 0$ and since $t_2, t_3 \geq 0$, (2.29) holds trivially. It remains to consider three signatures $\sigma =$ ("+", "+", "∘"), $\sigma =$ ("+", "∘", "+") and $\sigma =$ ("+", "∘", "−") where "∘" denotes a missing edge (which by our assumption above is a positive edge).

First, assume that $\sigma = (\text{“+”}, \text{“+”}, \text{“}\circ\text{”})$. By Claim 2.3.3, $t_1 = A(1-y_2)x_1 - (1-y_2) = -e^{-Ax_2}(1-Ax_1)$ and $t_2 = A(1-y_1)x_2 - (1-y_2) = e^{-Ax_1}(Ax_2-1)$. Since e_3 is missing, $w_3 = 0$. We have, $w_1t_1 + w_2t_2 + w_3t_3 \geq t_1 + \alpha t_2$ (here we used that $t_1 \leq 0$ and $t_2 \geq 0$). So it suffices to prove that $t_1 + \alpha t_2 > 0$ or, equivalently, $e^{Ax_2}(\alpha t_2 + t_1) \geq 0$. Using that $e^{A(x_2-x_1)} \geq 1/\alpha$ and $x_2 \geq \frac{A-1}{2A} - x_1$, we get

$$e^{Ax_2}(\alpha t_2 + t_1) = \alpha e^{A(x_2-x_1)}(Ax_2-1) - (1-Ax_1) \geq \alpha \cdot \frac{1}{\alpha} \cdot \Big(A\big(\frac{A-1}{2A} - x_1\big) - 1\Big) + Ax_1 - 1 = \frac{A-5}{2} > 0,$$

as required.

Now, assume that $\sigma = (\text{“+”}, \text{“}\circ\text{”}, \text{“+”})$. Now we have $t_1 = -e^{-Ax_2}(1-Ax_1)$ (as before) and

$$t_3 = A(1-y_1)x_3 - (y_2 - y_1) = Ae^{-Ax_1}x_3 - (e^{-Ax_1} - e^{-Ax_2}) = e^{-Ax_1}(Ax_3 - 1) + e^{-Ax_2}.$$

We prove that $t_1 + \alpha t_3 \geq 0$ or, equivalently, $e^{Ax_2}(\alpha t_3 + t_1) \geq 0$. Using that $e^{A(x_2-x_1)} \geq 1/\alpha$ and $x_3 \geq \frac{A-1}{2A}$, we get

$$\begin{aligned} e^{Ax_2}(\alpha t_3 + t_1) &= \alpha\big(e^{A(x_2-x_1)}(Ax_3 - 1) + 1\big) - (1 - Ax_1) \\ &\geq (Ax_3 - 1) + \alpha - (1 - Ax_1) > Ax_3 - 2 \geq \frac{A-1}{2} - 2 \geq 0, \end{aligned}$$

as required.

Finally, assume that $\sigma = (\text{“+”}, \text{“}\circ\text{”}, \text{“}-\text{”})$. Now we have $t_1 = -e^{-Ax_2}(1-Ax_1)$ (as before) and $t_3 = A(1-y_1)(1-x_3) - (1-y_2) = Ae^{-Ax_1}(1-x_3) - e^{-Ax_2}$. As in the previous

case, we prove that $e^{Ax_2}(\alpha t_3 + t_1) \geq 0$. We have,

$$e^{Ax_2}(\alpha t_3 + t_1) = \alpha\big(Ae^{A(x_2-x_1)}(1-x_3)-1\big)-(1-Ax_1) \geq \underbrace{\alpha\big(Ae^{A(x_2-x_1)}(1-x_1-x_2)-1\big)-(1-Ax_1)}_{F(x_1,x_2)}.$$

Denote the expression on the right by $F(x_1, x_2)$. We now show that for a fixed x_1, $F(x_1, x_2)$ is an increasing function of x_2 when $x_2 \in [\frac{A-1}{2A} - x_1, \frac{A-1}{2A})$. Indeed, we have

$$\begin{aligned}\frac{\partial F(x_1,x_2)}{\partial x_2} &= \alpha A e^{A(x_2-x_1)}\big(A(1-x_1-x_2)-1\big) \geq \alpha A e^{A(x_2-x_1)}\Big(A\Big(1-\frac{1}{A}-\frac{A-1}{2A}\Big)-1\Big)\\ &= \alpha A e^{A(x_2-x_1)} \cdot \frac{A-3}{2} > 0.\end{aligned}$$

We conclude that

$$\begin{aligned}F(x_1,x_2) \geq F\left(x_1, \frac{A-1}{2A} - x_1\right) &= \Big(\alpha\big(Ae^{A(\tilde{x}_2-x_1)}(1-x_1-\tilde{x}_2)-1\big)-(1-Ax_1)\Big)\Big|_{\tilde{x}_2=\frac{A-1}{2A}-x_1}\\ &\geq \alpha \cdot A \cdot \frac{1}{\alpha} \cdot \left(1-\frac{A-1}{2A}\right) - \alpha - (1-Ax_1) = \frac{A+1}{2} - \alpha - 1 + Ax_1 \geq \frac{A+1}{2} - 2 > 0.\end{aligned}$$

This concludes the case analysis and the proof of Theorem 1.4.2. □

2.6 Integrality Gap

In this section, we give a $\Theta(\log 1/\alpha)$ integrality gap example for the LP relaxation presented in Section 2.2.1. Notice that in the example each positive edge has a weight of $\boldsymbol{w}^+$ and each negative edge has a weight of $\boldsymbol{w}^-$ with $\boldsymbol{w}^+ \geq \boldsymbol{w}^-$.

Proof of Theorem 1.4.3. Consider a 3-regular expander $G = (V, E)$ on $n = \Theta((\alpha^2 \log^2 \alpha)^{-1})$ vertices. We say that two vertices u and v are similar if $(u, v) \in E$; otherwise u and v are dissimilar. That is, the set of positive edges E^+ is E and the set of negative edges E^- is $V \times V \setminus E$. Let $\boldsymbol{w}^+ = 1$ and $\boldsymbol{w}^- = \alpha$.

Lemma 2.6.1. *The integrality gap of the Correlation Clustering instance $G_{cc} = (V, E^+, E^-)$ described above is $\Theta(\log 1/\alpha)$.*

Proof. Let $d(u, v)$ be the shortest path distance in G. Let $\varepsilon = 2/\log_3 n$. We define a feasible metric LP solution as follows: $x_{uv} = \min(\varepsilon d(u, v), 1)$.

Let LP^+ be the LP cost of positive edges, and LP^- be the LP cost of negative edges. The LP cost of every positive edge is ε since $d(u, v) = 1$ for $(u, v) \in E$. There are $3n/2$ positive edges in G_{cc}. Thus, $LP_+ < 3n/\log_3 n$. We now estimate LP^-. For every vertex u, the number of vertices v at distance less than t is upper bounded by 3^t because G is a 3-regular graph. Thus, the number of vertices v at distance less than $1/2 \log_3 n$ is upper bounded by $\sqrt{n}$. Observe that the LP cost of a negative edge (u, v) (which is equal to $\alpha(1 - x_{uv})$) is positive if and only if $d(u, v) < 1/2 \log_3 n$. Therefore, the number of negative edges with a positive LP cost incident on any vertex u is at most $\sqrt{n}$. Consequently, the LP cost of all negative edges is upper bounded by $\alpha n^{\frac{3}{2}} = \Theta(n/\log 1/\alpha)$. Hence,

$$LP \leq \Theta(n/\log 1/\alpha) + 3n/\log_3 n = \Theta(n/\log 1/\alpha).$$

Here, we used that $\log n = \Theta(\log 1/\alpha)$.

We now lower bound the cost of the optimal (integral) solution. Consider an optimal solution. There are two possible cases.

1. No cluster contains 90% of the vertices. Then a constant fraction of positive edges in the expander G are cut and, therefore, the cost of the optimal clustering is at least $\Theta(n)$.

2. One of the clusters contains at least 90% of all vertices. Then all negative edges in that cluster are in disagreement with the clustering. There are at least $\binom{0.9n}{2} - m = \Theta(n^2)$ such edges. Their cost is at least $\Omega(\alpha n^2)$.

We conclude that the cost of the optimal solution is at least $\Theta(n)$ and, thus, the integrality gap is $\Theta(\log(1/\alpha))$. □

We note that in this example $\log(1/\alpha) = \Theta(\log n)$. However, it is easy to construct an integrality gap example where $\log(1/\alpha) \ll \Theta(\log n)$. To do so, we pick the integrality gap example constructed above and create $k \gg n$ disjoint copies of it. To make the graph complete, we add negative edges with (fractional) LP value equal to 1 to connect each copy to every other copy of the graph. The new graph has $kn \gg n$ vertices. However, the integrality gap remains the same, $\Theta(\log 1/\alpha)$. □

Now we give a $\Theta(\log 1/\alpha)$ integrality gap example when G is a complete bipartite graph.

Proof of Theorem 1.4.4. The proof is very similar to that of Theorem 1.4.3. We start with a 3-regular *bipartite* expander $G = (L, R, E)$ on $n = \Theta((\alpha^2 \log^2 \alpha)^{-1})$ vertices (e.g., we can use a 3-regular bipartite Ramanujan expander constructed by Marcus, Spielman, and Srivastava [36]). Then we define a correlation clustering instance as follows: $G_{cc} = (L, R, E^+, E^-)$ where $E^+ = E$ and $E^- = (L \times R) \setminus E$; let $\boldsymbol{w}^+ = 1$ and $\boldsymbol{w}^- = \alpha$. The proof of Lemma 2.6.1 can be applied to G_{cc}; we only need to note that if a cluster contains at least 90% of the vertices, then there are at least $\Theta(n^2)$ edges of G_{cc} between vertices in the cluster. It follows that the integrality gap is $\Omega(\log(1/\alpha))$. □

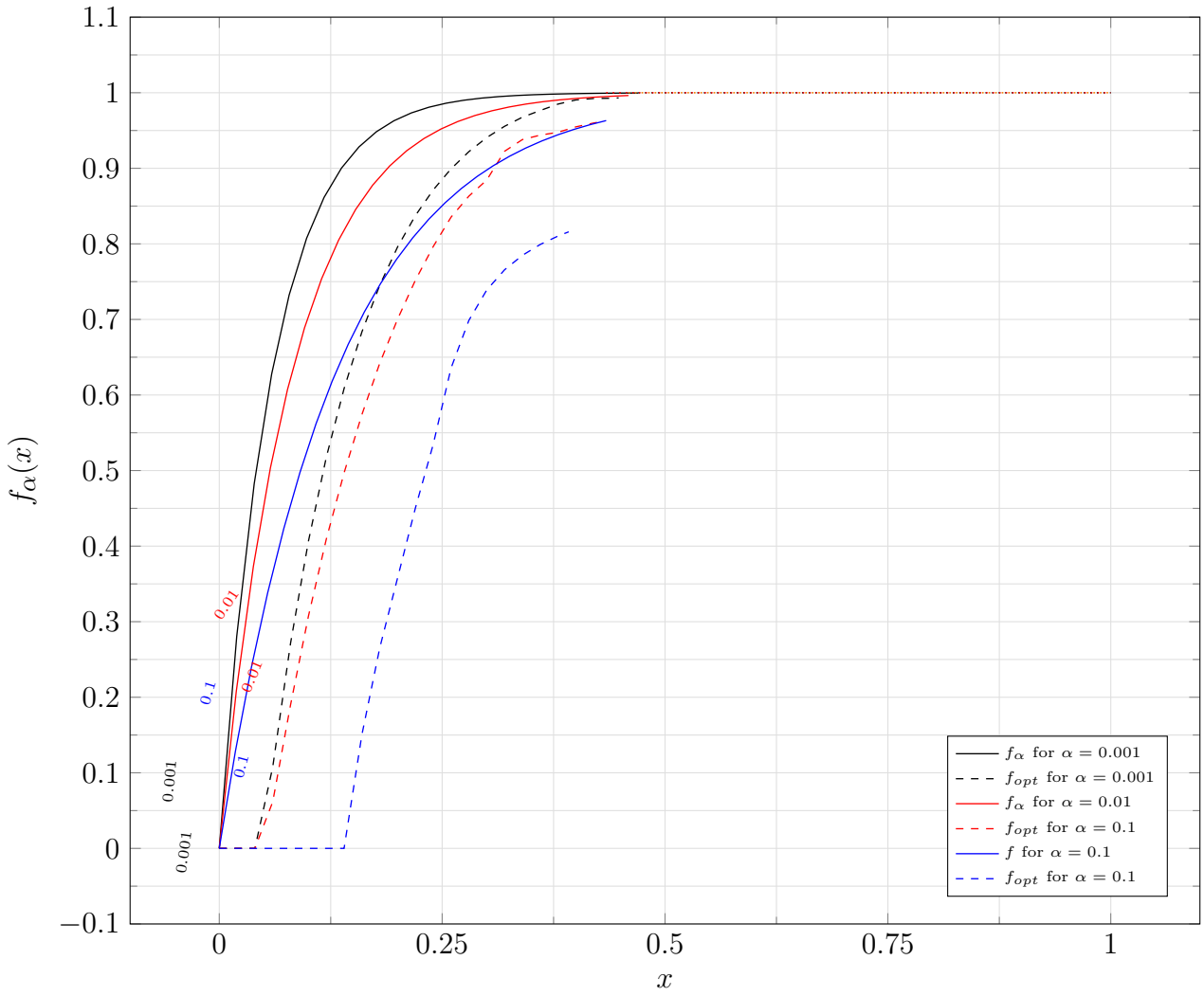

Figure 2.2: This plot shows functions $f_\alpha(x)$ used in the proof of Theorem 1.4.1 for $\alpha \in \{0.001, 0.01, 0.1\}$. Additionally, it shows optimal functions $f_{opt}(x)$ (see Section 2.4 for details). Note that every function $f_\alpha(x)$, including $f_{opt}(x)$, has a discontinuity at point $\tau = 1/2 - 1/2A$; for $x \geq \tau$, $f_\alpha(x) = 1$.

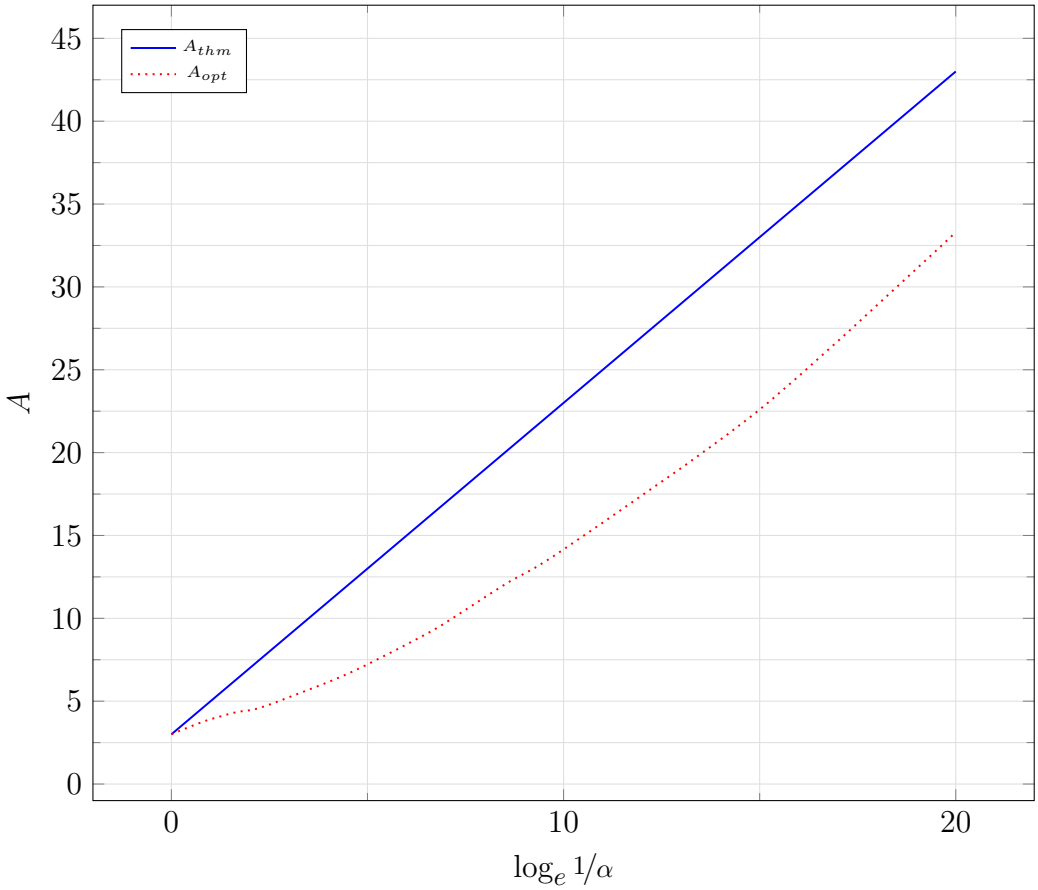

Figure 2.3: Plots of approximation factors A_{thm} and A_{opt}.

CHAPTER 3

APPROXIMATION ALGORITHM FOR THE ℓ_P OBJECTIVE

In this chapter, we study the ℓ_p objective in the Correlation Clustering with Asymmetric Classification Errors model. In Section 3.1, we describe the convex relaxation that we will use in our algorithm for Correlation Clustering. In Section 3.2, we introduce a novel technique for partitioning metric spaces. This forms the main technical basis for our algorithm for Correlation Clustering. In Section 3.3, we prove our second main result, Theorem 1.4.5. In Section 3.4, we describe our metric space partitioning scheme and give a proof overview of its correctness. In sections 3.5, 3.6, 3.6.3 and 4.2, we formally prove the correctness of our partitioning scheme, Theorem 3.2.1. In Section 3.7, we prove our integrality gap result, Theorem 1.4.8.

3.1 Convex Relaxation

Our algorithm for minimizing local objectives is based on rounding the optimal solution to a suitable convex program (Figure 3.1). This convex program is similar to the relaxations used in Charikar et al. [17] and Kalhan et al. [33]. In this convex program, we have a variable x_{uv} for every pair of vertices $u, v \in V$. The variable x_{uv} captures the distance between u and v in the "multicut metric". In the integral solution, $x_{uv} = 0$ if u and v are in the same partition and $x_{uv} = 1$ if u and v are in different partitions. In order to enforce that the partitioning is consistent, we add triangle inequality constraints between all triplets of vertices (P2). We also require that distance x_{uv} is symmetric (P3).

For every vertex $u \in V$, we use the variable y_u to denote the total weight of violated edges incident on u (P1). The objective of the convex program is thus to minimize $\|y\|_p$ – the ℓ_p norm of the vector y. Notice that each constraint in the convex program is linear, and the objective function $\|\cdot\|_p : \mathbb{R}^n \to \mathbb{R}$ is convex (by the Minkowski inequality).

$$
\begin{array}{lll}
\text{minimize} & \|y\|_p & \text{(P)} \\
\text{subject to} & y_u = \sum_{v:(u,v)\in E^+} w_{uv}x_{uv} + \sum_{v:(u,v)\in E^-} w_{uv}(1-x_{uv}) \quad \text{for all } u \in V & \text{(P1)} \\
& x_{v_1v_2} + x_{v_2v_3} \geq x_{v_1v_3} \quad \text{for all } v_1, v_2, v_3 \in V & \text{(P2)} \\
& x_{uv} = x_{vu} \quad \text{for all } u, v \in V & \text{(P3)} \\
& x_{uv} \in [0,1] \quad \text{for all } u, v \in V & \text{(P4)}
\end{array}
$$

Figure 3.1: Convex relaxation for Correlation Clustering with min ℓ_p objective for $p \geq 1$ or $p = \infty$.

What remains to be shown is that the relaxation presented in Figure 3.1 is valid. To this end, consider any partition $\mathcal{P} = (P_1, P_2, \ldots, P_k)$ of the set of vertices V. For every pair of vertices u, v, if u and v lie in the same partition, we assign the corresponding variable x_{uv} a value of 0, else we assign it a value of 1. Note that such an assignment satisfies the triangle inequality (P2). Variable y_u thus captures the total weight of violated edges incident on u; every similar edge (u, v) incident on u that crosses a partition contributes $w_{uv} \cdot x_{uv} = w_{uv}$ to y_u, and every dissimilar edge present within a cluster contributes $w_{uv} \cdot (1 - x_{uv}) = w_{uv}$ to y_u. Thus, y_u is equal to $\text{dis}_u(\mathcal{P}, E^+, E^-)$. Hence, an integral convex program solution defined in such a manner is feasible and has the same cost as the partitioning. It is possible, however, that the cost of the optimal fractional solution is less than the cost of the optimal integral solution, and hence the convex program in Figure 3.1 is a relaxation to our problem. We note that our relaxation is simpler than the relaxation used in Kalhan et al. [33]. The additional variables in their convex program are not needed in our case because all edge weights belong to the interval $[\alpha\mathbf{w}, \mathbf{w}]$.

3.2 A New Technique for Partitioning Metric Spaces

We will use the following notation: Given expressions X and Y, we write $X \lesssim Y$ if $X \leq C \cdot Y$ for some constant $C > 0$ (that is, $X = O(Y)$). We define $\gtrsim$ similarly. Furthermore, let

$X^+ = 0$ if $X < 0$ and $X^+ = X$ if $X \geq 0$. We use $\mathrm{Ball}(v, l) = \{u : d(u, v) \leq l\}$ to denote the set of vertices at a distance of at most l from v.

In this section, we describe our main technical tool – a novel probabilistic scheme for partitioning metric spaces which may be of independent interest. This partitioning scheme forms the basis of our algorithm (Algorithm 3) for Correlation Clustering. We begin by stating this technical result.

Theorem 3.2.1. *For every $q \geq 1$ there exists a $\beta_q^* = \Theta\big(\frac{1}{q \ln(q+1)}\big) < 1$ such that the following holds. Consider a finite metric space (X, d). Fix two positive numbers r and R such that $\beta = r/R \leq \beta_q^*$. Let $D_\beta = 2(q+1)\ln 1/\beta$. Then, there exists a probabilistic partitioning $\mathcal{P}$ satisfying properties (1), (2), and (3):*

(1) $\mathrm{diam}(P) \leq 2R$ *for every $P \in \mathcal{P}$ (always);*

(2) For every point u in X, the following bound holds:

$$\sum_{v \in \mathrm{Ball}(u,R)} \left(\Pr\left\{\mathcal{P}(u) \neq \mathcal{P}(v)\right\} - D_\beta \frac{d(u,v)}{R} \right)^+ \lesssim \beta^q \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R},$$

where $\mathcal{P}(u)$ denotes the partition of $\mathcal{P}$ that contains u.

(3) Moreover, for every u in X, we always have,

$$\sum_{v \in \mathrm{Ball}(u,r)} \mathbb{1}\left\{\mathcal{P}(u) \neq \mathcal{P}(v)\right\} \lesssim \beta \cdot D_\beta^2 \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}.$$

The partitioning we construct in Theorem 3.2.1 resembles a $2D$–separating $2R$-bounded stochastic decomposition of a metric space [8, 15, 27]. Recall that a $2D$–separating $2R$-bounded stochastic decomposition satisfies property (1) of Theorem 3.2.1 and the $2D$-

separating condition: for every $u, v \in X$,

$$\Pr\left\{\mathcal{P}(u) \neq \mathcal{P}(v)\right\} - D\frac{d(u,v)}{R} \leq 0. \tag{3.1}$$

At a very high level, the goals of our partitioning and the $2D$–separating $2R$-bounded stochastic decomposition are similar: decompose a metric space in clusters of diameter at most $2R$ so that nearby points lie in the same cluster with high enough probability. However, the specific conditions are quite different. Loosely speaking, property (2) of Theorem 3.2.1 says that the decomposition satisfies (3.1) with $D = D_\beta$ on average up to an additive error term of $O(\beta^q) \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}$. Crucially, property (3) provides an analogous guarantee not only in expectation, but also in the worst case (which a $2D$-separating decomposition does not satisfy).

Property (3) plays a key role in proving our main result, Theorem 1.4.5. For the standard objective function for Correlation Clustering (minimizing the ℓ_1 norm of the disagreements vector), properties (1) and (2) are sufficient since an upper bound on the expected weight of disagreements on a single vertex implies an upper bound on the expected weight of the total disagreements. The situation gets trickier when we consider minimizing arbitrary ℓ_p ($p > 1$) norms of the disagreements vector. For instance, having an upper bound on the expected weight of disagreements on a single vertex does not necessarily translate to an upper bound on the expected weight of disagreements on a worst vertex (ℓ_∞ norm). We overcome this nonlinear nature of the problem for higher values of p by using the deterministic (worst-case) guarantee given by property (3) of Theorem 3.2.1.

Also note that coefficients D_β and β do not depend on the size $|X|$ of the metric space (in our algorithm, they will only depend on α, which is defined as the ratio of the smallest edge weight to the largest positive edge weight). However, the optimal value of D in the $2D$-separating condition is $\Theta(\log |X|)$.

3.3 Correlation Clustering via Metric Partitioning

In this section, we will prove our main theorem, Theorem 1.4.5. Our algorithm (Algorithm 3) for minimizing local objectives for Correlation Clustering with Asymmetric Classification Errors begins by solving the convex relaxation in Figure 3.1 to obtain a solution $\{x_{uv}\}_{u,v \in V}$. It then defines a metric $d(\cdot, \cdot)$ on V by setting distances $d(u, v) = x_{uv}$.

We let $q = 2$. Let α^* be the solution of equation $3\sqrt{\alpha^*}/\ln 1/\alpha^* = \beta_2^*$ (note that α^* is an absolute constant). We assume that $\alpha \leq \alpha^*$. If $\alpha > \alpha^*$, we just redefine α as α^* (this will increase the approximation ratio only by a constant factor). We set $r = \sqrt{\alpha}/\ln 1/\alpha$ and $R = 1/3$. Note that $r/R \leq \beta_2^* < 1$.

At this point, the algorithm makes use of our key technical contribution – a new probabilistic scheme for partitioning metric spaces (Algorithm 4) – and outputs the partitioning thus obtained. Please refer to Algorithm 3 for a summary.

To show that $\mathcal{P}$ has the desired approximation ratio in Theorem 1.4.5, we bound the weight of disagreements at every vertex $u \in V$ with respect to $\mathcal{P}$. To this end, we show that two useful quantities, the total weight of disagreements at u and the expected weight of disagreements at u can be bounded in terms of y_u, the cost paid by the convex program for vertex u. In Theorem 3.3.1, we make use of the properties of $\mathcal{P}$ given by Theorem 3.2.1 to get a bound on these two quantities for each vertex $u \in V$. Then, in Section 3.3.1, we use the bounds from Theorem 3.3.1 to complete the proof of Theorem 1.4.5: we show that if the total cost of disagreements and the expected cost of disagreements with respect to $\mathcal{P}$ are bounded for every $u \in V$, then the partitioning $\mathcal{P}$ achieves the desired approximation ratio in Theorem 1.4.5. We remind the reader that given a partitioning $\mathcal{P}$ of the vertex set and a vertex $u \in V$, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$ denotes the weight of edges incident on u that are in disagreement with respect to $\mathcal{P}$. Moreover, y_u denotes the convex programming (CP) cost of the vertex u.

Define $A_1 = \ln 1/\alpha$ and $\mathrm{A}_\infty = \ln(\frac{1}{\alpha})/\sqrt{\alpha} = 1/r$. Our analysis focuses on bounding two

Algorithm 3 Correlation Clustering Algorithm

Input: $G = (V, E^+, E^-, \mathbf{w}, \alpha)$, $\{x_{uv}\}_{u,v \in V}$.
Define a metric d on V such that $d(u, v) = x_{uv}$ for all $u, v \in V$.
Define $r = (\sqrt{\alpha}/\ln 1/\alpha), R = 1/3$, $q = 2$.
$\mathcal{P}$ = Metric Space Partitioning Scheme(V, d, r, R, q).
Output $\mathcal{P}$.

key quantities related to a vertex $u \in V$. The first quantity, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$, is the total weight of edges incident on u that are in disagreement with $\mathcal{P}$. We show that this quantity can be charged to the CP cost of u and is at most $\mathrm{A}_\infty \cdot y_u$. We then get a stronger bound for our second quantity of interest, $\mathbb{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)]$, the expected cost of a vertex u. In particular, we show that $\mathbb{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \leq A_1 \cdot y_u$.

Theorem 3.3.1. *Given an instance of Correlation Clustering with Asymmetric Classification Errors (Definition 2), Algorithm 3 outputs a partitioning $\mathcal{P}$ of the vertex set such that the following holds for every vertex $u \in V$:*

(a) $\mathrm{dis}_u(\mathcal{P}, E^+, E^-) \lesssim A_\infty \cdot y_u$*;*

(b) $\mathbb{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \lesssim A_1 \cdot y_u$*,*

where $A_1 = \ln(1/\alpha)$ *and* $\mathrm{A}_\infty = \ln(\frac{1}{\alpha})/\sqrt{\alpha}$.

Proof. Without loss of generality we assume that the scaling parameter $\mathbf{w}$ is 1. Thus, for every positive edge $e^+ \in E^+$, $w_{e^+} \in [\alpha, 1]$, while for every negative edge $e^- \in E^-, w_{e^-} \geq \alpha$. Write the formula for $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$ for a given vertex $u \in V$,

$$\mathrm{dis}_u(\mathcal{P}, E^+, E^-) = \sum_{(u,v) \in E^+} w_{uv} \cdot \mathbb{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} + \sum_{(u,v) \in E^-} w_{uv} \cdot \mathbb{1}\{\mathcal{P}(u) = \mathcal{P}(v)\}.$$

Let $E^{\geq r}$ be the set of positive edges (v, w) in E^+ with $x_{vw} \geq r$. Observe that

$$\mathrm{dis}_u(\mathcal{P}, E^+, E^-) = \quad \mathrm{dis}_u(\mathcal{P}, \varnothing, E^-) + \mathrm{dis}_u(\mathcal{P}, E^{\geq r}, \varnothing) + \mathrm{dis}_u(\mathcal{P}, E^+ \setminus E^{\geq r}, \varnothing). \qquad (3.2)$$

Recall that $\beta = r/R = 3\sqrt{\alpha}/\ln 1/\alpha$, $q = 2$, and $D_\beta = \Theta(\ln 1/\beta) = \Theta(\ln 1/\alpha)$. From Theorem 3.2.1, part (a), we know that the diameter of each partition P in $\mathcal{P}$ is at most $2R$. For any negative edge to be in disagreement, both its endpoints must lie in the same partition. Thus, the length x_{uv} for any such edge $(u, v) \in E^-$ is at most $2R$, and hence its CP contribution is at most $(1 - 2R) = 1/3$. Hence,

$$\mathrm{dis}_u(\mathcal{P}, \varnothing, E^-) = \sum_{(u,v)\in E^-} w_{uv} \mathbb{1}\{\mathcal{P}(u) = \mathcal{P}(v)\} \leq 3y_u.$$

Then,

$$\mathrm{dis}_u(\mathcal{P}, E^{\geq r}, \varnothing) \leq |\{v : (u, v) \in E^{\geq r}\}| \leq \frac{y_u}{r} = \mathrm{A}_\infty\, y_u.$$

To complete the proof of Theorem 3.3.1, part (a) we write:

$$\begin{aligned}\mathrm{dis}_u(\mathcal{P}, E^+ \setminus E^{\geq r}, \varnothing) &= \sum_{v\in \mathrm{Ball}(u,r)} w_{uv} \cdot \mathbb{1}\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\} \\ &\leq \sum_{v\in \mathrm{Ball}(u,r)} \mathbb{1}\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\}.\end{aligned}$$

The inequality above holds because the weight of each positive edge is at most 1. Next, using the bound for $\sum_{v\in \mathrm{Ball}(u,r)} \mathbb{1}\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\}$ from Theorem 3.2.1 part (c), we get,

$$\sum_{v\in\mathrm{Ball}(u,r)} \mathbb{1}\{\mathcal{P}(u)\neq\mathcal{P}(v)\} \lesssim \beta\cdot D_\beta^2 \sum_{v\in\mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}$$
$$\lesssim \frac{\sqrt{\alpha}}{\ln(1/\alpha)}\cdot(\ln^2(1/\alpha)) \sum_{v\in\mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}$$
$$\lesssim \frac{\sqrt{\alpha}}{\ln(1/\alpha)}\cdot\ln^2(1/\alpha)\cdot\frac{y_u}{\alpha} = A_\infty\cdot y_u,$$

where the last inequality follows from the fact that each positive edge weight is at least α. Thus, from (3.2) it follows:

$$\mathrm{dis}_u(\mathcal{P},E^+,E^-) \lesssim \mathrm{A}_\infty\cdot y_u.$$

We now prove Theorem 3.3.1, part (b). We separately consider short and long positive edges. Let $E^{\le R}$ be the set of positive edges $(v,w)\in E^+$ with $x_{vw}\le R$. Note that

$$y_u \ge \sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\min(d(u,v),1-d(u,v)) \tag{3.3}$$
$$= \sum_{v\in\mathrm{Ball}(u,R)} w_{uv}d(u,v) = \frac{1}{3}\sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\frac{d(u,v)}{R}.$$

Therefore, we have

$$\mathbb{E}[\mathrm{dis}_u(\mathcal{P},E^{\le R},\varnothing) - 3D_\beta\cdot y_u] \le \mathbb{E}\Big[\sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\cdot\mathbb{1}\{\mathcal{P}(u)\neq\mathcal{P}(v)\} - D_\beta\sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\frac{d(u,v)}{R}\Big]$$
$$= \sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\Big(\Pr\{\mathcal{P}(u)\neq\mathcal{P}(v)\} - D_\beta\frac{d(u,v)}{R}\Big)$$
$$\le \sum_{v\in\mathrm{Ball}(u,R)} w_{uv}\Big(\Pr\{\mathcal{P}(u)\neq\mathcal{P}(v)\} - D_\beta\frac{d(u,v)}{R}\Big)^+.$$

Since all edges (u,v) in $E^{\le R}$ are positive, we have $w_{uv} \le 1$. Consequently,

$$\mathbb{E}[\operatorname{dis}_u(\mathcal{P}, E^{\le R}, \varnothing) - 3D_\beta \cdot y_u] \le \sum_{\substack{v \in \mathrm{Ball}(u,R) \\ \text{s.t. } (u,v)\in E^+}} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R}\Big)^+.$$

We bound the right hand side using property (2) of Theorem 3.2.1:

$$\begin{aligned}
\sum_{v\in \mathrm{Ball}(u,R)} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R}\Big)^+ &\lesssim \beta^2 \sum_{v\in\mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \\
&\lesssim \frac{\alpha}{\ln^2(1/\alpha)} \sum_{v\in\mathrm{Ball}(u,2R)} d(u,v) \\
&\le \frac{1}{\ln^2(1/\alpha)} \sum_{v\in\mathrm{Ball}(u,2R)} w_{uv} \cdot 2\min(d(u,v), 1-d(u,v)) \\
&\le \frac{2}{\ln^2(1/\alpha)} \cdot y_u.
\end{aligned}$$

Here, we used that $w_{uv} \ge \alpha$ and $d(u,v) \le 2(1-d(u,v))$ for $v \in \mathrm{Ball}(u,2R)$. Thus,

$$\mathbb{E}[\operatorname{dis}_u(\mathcal{P}, E^{\le R}, \varnothing)] \lesssim \big(\ln(1/\alpha) + \frac{1}{\ln^2(1/\alpha)}\big) y_u \lesssim A_1 \cdot y_u.$$

Furthermore, $\operatorname{dis}_u(\mathcal{P}, E^+ \setminus E^{\le R}, \varnothing) \le \frac{1}{R} \cdot y_u \le A_1 \cdot y_u$. Therefore, from (3.2) it follows that

$$\mathbb{E}[\operatorname{dis}_u(\mathcal{P}, E^+, E^-)] \lesssim A_1 y_u.$$

□

We now use Theorem 3.3.1 to prove Theorem 1.4.5.

3.3.1 Proof of Theorem 1.4.5

In this section, we show that the partitioning $\mathcal{P}$ output by Algorithm 3 achieves the desired approximation ratio – thereby proving our main theorem, Theorem 1.4.5. To show this, we will use the fact that $\mathcal{P}$ satisfies the properties in Theorem 3.3.1.

Proof of Theorem 1.4.5. If $p = \infty$, then we get an $O(A_\infty) = O((1/\alpha)^{1/2} \ln 1/\alpha)$ approximation by Theorem 3.3.1, item (a), as desired. So we assume that $p < \infty$ below. Given the guarantees from Theorem 3.3.1, we observe,

$$\begin{aligned}
\mathbb{E}\left[\sum_{u\in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p\right] &= \sum_{u\in V} \mathbb{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)^{p-1} \cdot \mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \\
&\lesssim \sum_{u\in V} \mathbb{E}\left[(\mathrm{A}_\infty \cdot y_u)^{p-1} \cdot \mathrm{dis}_u(\mathcal{P}, E^+, E^-)\right] \\
&= \sum_{u\in V} (\mathrm{A}_\infty \cdot y_u)^{p-1} \mathbb{E}\left[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)\right] \\
&\lesssim \sum_{u\in V} (\mathrm{A}_\infty \cdot y_u)^{p-1} \cdot A_1 \cdot y_u = \sum_{u\in V} A^p \cdot y_u^p,
\end{aligned}$$

where $A = \left(\mathrm{A}_\infty^{p-1} \cdot A_1\right)^{\frac{1}{p}}$. Note that the desired approximation factor is $O(A)$. From Jensen's inequality, it follows that

$$\begin{aligned}
\mathbb{E}\left[\Big(\sum_{u\in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p\Big)^{\frac{1}{p}}\right] &\le \left(\mathbb{E}\left[\sum_{u\in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p\right]\right)^{\frac{1}{p}} \\
&\lesssim \left(\sum_{u\in V} A^p \cdot y_u^p\right)^{\frac{1}{p}} = A \cdot \|y\|_p.
\end{aligned}$$

This finishes the proof. □

3.4 Overview of Metric Partitioning Scheme

In this section we describe our partitioning scheme and give a proof overview of Theorem 3.2.1. A pseudocode for this partitioning scheme is given in Algorithm 4. More specifically, in Section 3.4.1 we reduce the problem to choosing a random set of particular interest as stated in Theorem 3.4.1. In Section 3.4.2 we describe an algorithm for choosing such a random set and give a proof overview of its correctness. The pseudocode for choosing a random set is given in Algorithm 5.

3.4.1 Iterative Clustering

Given a metric space (X, d), our partitioning scheme uses an iterative algorithm – Algorithm 4 to obtain $\mathcal{P}$. Let X_t denote the set of not-yet clustered vertices at the start of iteration t of Algorithm 4. At step t, the algorithm finds and outputs random set $P_t \subseteq X_t$. It then updates the set of not-yet clustered vertices ($X_{t+1} = X_t \setminus P_t$), and repeats this step until all vertices are clustered. Algorithm 4 makes use of the following theorem in each iteration to find the random set P_t.

We need the following notation to state the theorem. Let $\delta_P(u, v)$ be the cut metric induced by the set P: $\delta_P(u, v) = 1$ if $u \in P$ and $v \notin P$ or $u \notin P$ and $v \in P$; $\delta_P(u, v) = 0$ if $u \in P$ and $v \in P$ or $u \notin P$ and $v \notin P$. Also, let $\vee_P(u, v)$ be the indicator of the event $u \in P$ or $v \in P$ or both u and v are in P. We denote $[k] = \{1, 2, \dots k\}$.

Theorem 3.4.1. *For every $q \geq 1$ there exists a $\beta_q^* = \Theta\big(\frac{1}{q \ln(q+1)}\big) < 1$ such that the following holds. Consider a finite metric space (X, d). Fix two positive numbers r and R such that $\beta = r/R \leq \beta_q^*$. Let $D_\beta = 2(q+1) \ln 1/\beta$. Then, there exists an algorithm for finding a random set P satisfying properties (a), (b), and (c):*

(a) $\operatorname{diam}(P) \leq 2R$ *(always);*

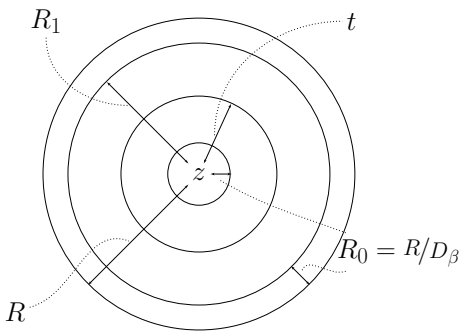

Figure 3.2: Balls with Different Radii
$R > r > 0$, $q \geq 1$, $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$, $R_0 = R/D_\beta$, $R_1 = R - R_0$.

(b) For every point u in X, the following bound holds:

$$\sum_{v \in \mathrm{Ball}(u,R)} \Big(\Pr\big\{\delta_P(u,v) = 1\big\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_P(u,v) = 1\} \Big)^+ \lesssim \beta^q \cdot \mathbb{E}\left[\sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_P(u,v) \right].$$

(c) Moreover, for every u in X, we always have

$$\sum_{v \in \mathrm{Ball}(u,r)} \delta_P(u,v) \lesssim \beta \cdot D_\beta^2 \cdot \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_P(u,v).$$

Informally, Theorem 3.4.1 is a "single-cluster" version of Theorem 3.2.1, and there is a one-to-one correspondence between their properties. In Section 3.5, we show that Theorem 3.2.1 holds for $\mathcal{P}$ if we assume that each partition $P \in \mathcal{P}$ satisfies Theorem 3.4.1. Thus, to obtain Theorem 3.2.1, it remains to prove Theorem 3.4.1.

3.4.2 Selecting a Single Cluster

We will use the following definitions. Let r and R be positive numbers with $r < R$. Define $\beta = r/R \leq \beta_q^*$ and $D_\beta = 2(q+1)\ln 1/\beta$ where $q \geq 1$. Let $R_0 = R/D_\beta$ and $R_1 = R - R_0$. We let $\rho_q(\beta) = (1/\beta)^{q+1}$ (see Figure 3.2). We choose β_q^* so that $r < R_0 < R$ (see Section 3.6 for details).

Algorithm 4 Metric Space Partitioning Scheme

Input: Metric Space (X, d) and $r, R > 0$, $q \geq 1$.
Define $t = 0$, $X_1 = X$.
repeat
  $t = t + 1$.
  $P_t = \text{Cluster Select}(X_t, d, r, R, q)$.
  $X_{t+1} = X_t \setminus P_t$.
until $X_t = \varnothing$
Output $(P_1, P_2, \ldots, P_t)$.

Given a metric space (X, d) and parameters r and R, our procedure for finding a random set $P \subseteq X$ begins by finding a pivot point z with a densely populated neighborhood – namely, z is chosen such that a ball of radius R_0 around z contains the maximum number of points. More formally,

$$z = \arg\max_{u \in X} |\operatorname{Ball}(u, R_0)|. \tag{3.4}$$

We refer to this ball of small radius around z as the "core" of the cluster. Our choice of the pivot z is inspired by the papers by Charikar et al. [16], Puleo and Milenkovic [40], Charikar et al. [17]. We then consider a ball of large radius R_1 around the pivot z and examine the following two cases – "Heavy Ball" and "Light Ball". If this ball of large radius around z is sufficiently populated, that is, if the number of points in $\operatorname{Ball}(z, R_1)$ is at least $(1/\beta)^{q+1}$ times the number of points in the core, we call this case "Heavy Ball". In the case of Heavy Ball, we will show that $P = \operatorname{Ball}(z, R_1)$ (a ball around z of radius slightly less than R) satisfies the properties of Theorem 3.4.1. In the case of "Light Ball", the ball of large radius around z is not sufficiently populated. In this case, the algorithm finds a radius t ($t \leq R$) such that $P = \operatorname{Ball}(z, t)$ satisfies the properties of Theorem 3.4.1. In the following subsections we provide an overview of the proof for these two cases. A formal proof of Theorem 3.4.1 can be found in Section 3.6.

Algorithm 5 Cluster Select

Input: Metric space (X, d) and $r, R > 0$, $q \geq 1$
Define: $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$.
Define: $R_0 = R/D_\beta$, $R_1 = R - R_0$, $\rho_q(\beta) = (1/\beta)^{q+1}$.
Select $z = \arg\max_{u \in X} |\operatorname{Ball}(u, R_0)|$.
if $|\operatorname{Ball}(z, R_1)| \geq \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$ **then**
  Set $P = \operatorname{Ball}(z, R_1)$.
else
  Consider S as stated in Definition 3.
  Consider π_S^{inv} as stated in Definition 4.
  Let F be the cumulative distribution function stated in Definition 5.
  Choose a random $x \in [0, R/2]$ according to F.
  Set $P = \operatorname{Ball}(z, \pi_S^{inv}(x))$.
end if
Output P.

Heavy Ball

The Heavy Ball P is a ball of radius R_1 around z which contains many points. As the diameter of P is $2R_1 < 2R$, it is easy to see that a Heavy Ball satisfies property (a) of Theorem 3.4.1. We now focus on showing that properties (b) and (c) hold for Heavy Ball. Observe as z was chosen according to (3.4), for every point $u \in X \setminus \{z\}$, u has a less populated neighborhood of radius R_0 than that of z. This combined with the fact that $\operatorname{Ball}(z, R_1)$ is heavy, implies that for every u, there are sufficiently many points in P at a distance of at least R_0 from u. Thus, for any point $u \in X$, we can expect the sum of distances between u and the points in P to be large. In fact, we show that the left hand sides of properties (b) and (c) can be charged to $\sum_{v \in P} \frac{d(u,v)}{R}$, the sum of distances between u and the points in P. For points u such that $d(z, u) \leq R$, $P \subseteq \operatorname{Ball}(u, 2R)$ and hence, $\sum_{v \in \operatorname{Ball}(u,2R)} \frac{d(u,v)}{R} \vee_P (u, v) \geq \sum_{v \in P} \frac{d(u,v)}{R}$. Thus, for every $u \in X$, we can charge the left hand sides of properties (b) and (c) to the quantity $\sum_{v \in \operatorname{Ball}(u,2R)} \frac{d(u,v)}{R} \vee_P (u, v)$. This allows us to conclude that a Heavy Ball satisfies Theorem 3.4.1.

Light Ball

In this subsection, we consider the case of $|\operatorname{Ball}(z, R_1)| < \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$, which we call Light Ball. In the case of Light Ball, we choose a random radius $t \in (0, R_1]$ and set $P = \operatorname{Ball}(z, t)$. Observe that property (a) of Theorem 3.4.1 holds trivially since the radius $t < R$.

Now consider property (c) of Theorem 3.4.1. Recall that for every point $u \in X$, property (c) gives a bound on the total number of points separated from u (by P) residing in a small ball $\operatorname{Ball}(u, r)$, i.e., $\sum_{v \in \operatorname{Ball}(u,r)} \delta_P(u, v)$. Note that property (c) gives a deterministic guarantee on P. Therefore, we choose a random radius $t \in (0, R_1]$ from the set of all radii for which property (c) of Theorem 3.4.1 holds. More specifically, we define the following set.

Definition 3. *Let S be the set of all radii s in $(3R_0, R_1]$ such that for every $u \in X$ set $P = \operatorname{Ball}(z, s)$ satisfies:*

$$\sum_{v \in \operatorname{Ball}(u,r)} \delta_P(u, v) \leq 25\beta \cdot D_\beta^2 \cdot \sum_{v \in \operatorname{Ball}(u,2R)} \frac{d(u, v)}{R} \cdot \vee_P(u, v). \tag{3.5}$$

The set S can be computed in polynomial time since the number of distinct clusters $P = \operatorname{Ball}(z, t)$ is upper bounded by the size of the metric space, $|X|$. By the same token, S is a finite union of disjoint intervals.

Now we show why we can expect the set S to be large. Consider $P = \operatorname{Ball}(z, s)$ such that $s \in S$. As S is computed according to Definition 3, it implies that the boundary of P is somewhat sparsely populated – as for every $u \in X$, it bounds the number of points within a small neighborhood of $\operatorname{Ball}(u, r)$ that are separated from u (note that $\sum_{v \in \operatorname{Ball}(u,r)} \delta_P(u, v)$ is trivially 0 for points u that are not close to the boundary of P). Since $\operatorname{Ball}(z, R_1)$ does not contain many points, the number of points in $\operatorname{Ball}(z, s')$ cannot grow too quickly as we increase the radius s' from 0 to R_1. This suggests that for many of such radii s', the ball $P = \operatorname{Ball}(z, s')$ has a sparsely populated boundary, and hence the set S should be large. In

fact, we use the above argument to show that the Lebesgue measure of the set S satisfies $\mu(S) \geq R/2$. This will allow us to define a continuous probability distribution on S.

What remains to be shown is that for a random radius $t \in S$, the set $P = \text{Ball}(z, t)$ satisfies property (b) of Theorem 3.4.1. For this purpose we define a measure preserving transformation π_S that maps an arbitrary measurable set S to the interval $[0, \mu(S)]$.

Definition 4. *Consider a measurable set $S \subset [0, R]$. Define function $\pi_S : [0, R] \to [0, \mu(S)]$ as follows $\pi_S(x) = \mu([0, x] \cap S)$. Also, for $y \in [0, \mu(S)]$, let*

$$\pi_S^{inv}(y) = \min\{x : \pi_S(x) = y\}.$$

Recall that the set S stated in Definition 3 is a finite union of disjoint intervals. In this case, what π_S does is simply pushing the intervals in S towards 0, and thus, allowing us to treat the set S as a single interval $[0, \mu(S)]$. For the rest of the proof overview, we assume that $S = [0, \mu(S)]$ and π_S is the identity. This simplifies the further analysis of Theorem 3.4.1 immensely.

Next, we define a cumulative distribution function F on $[0, R/2] \subseteq [0, \mu(S)]$:

Definition 5. *Let $F : [0, R/2] \to [0, 1]$ be a cumulative distribution function such that*

$$F(x) = \frac{1 - e^{-x/R_0}}{1 - e^{-R/2R_0}}. \tag{3.6}$$

We choose a random $x \in [0, R/2]$ according to F and set $P = \text{Ball}(z, \pi_S^{inv}(x))$ (see Algorithm 5). Since we assume in this proof overview that π_S is the identity, $P = \text{Ball}(z, x)$. Now, we show that the radius x chosen in such a manner guarantees that the cluster P satisfies property (b). Loosely speaking, the motivation behind our particular choice of cumulative distribution function F is the following: For two points $u, v \in X$, function F bounds the probability of u and v being separated by P, in terms of D_β times the probability that either u or v lies in P. Unfortunately, this bound does not hold for points u with

$d(z,u)$ close to $R/2$. However, the choice of parameters for function F in Definition 5 gives us two desired properties. Without loss of generality assume that $d(z,u) \leq d(z,v)$. Then, the probability that P separates the points u and v, $\Pr(\delta_P(u,v)) = \Pr(d(z,u) \leq x \leq d(z,v)) = F(d(z,v)) - F(d(z,u))$. Moreover, as $d(z,u) \leq d(z,v)$, the probability that either u or v lies in P, $\Pr(\vee_P(u,v)) = 1 - F(d(z,u))$. Thus, choosing F according to Definition 5 ensures:

- (Property I) $F(d(z,v)) - F(d(z,u))$ is bounded in terms of D_β times $1 - F(d(z,u))$ (Please see Claim 3.6.9 for a formal argument).
- (Property II) The probability that the cluster P includes points u such that $d(z,u) > R/2 - R_0$, is very small (please see Claim 3.6.8).

In fact, (Property II) of function F is the reason why we are able to guarantee that property (b) satisfies (3.1) only on average, with the error term coming from our inability to guarantee (3.1) for points on the boundary. We refer the reader to Section 3.6.5 for a formal proof. Thus, we conclude the case of Light Ball and show that it satisfies Theorem 3.4.1.

3.5 Proof of Theorem 3.2.1

In this section, we present the proof of our main technical result – Theorem 3.2.1 – an algorithm for partitioning a given metric space (X,d) into a number of clusters $\mathcal{P} = (P_1, \ldots, P_k)$ (where k is not fixed).

Recall our iterative process for obtaining this partitioning – Algorithm 4 – which makes use of Theorem 3.4.1 in each iteration to select a cluster from the set of not-yet clustered vertices.

The proof of Theorem 3.4.1 is presented in Section 3.6. We now present the proof of Theorem 3.2.1 assuming Theorem 3.4.1.

Proof of Theorem 3.2.1. Property (a) of Theorem 3.4.1 guarantees that $\mathrm{diam}(P_i) \leq 2R$ for every $i \in [k]$ and thus property (1) of Theorem 3.2.1 holds.

We now show that property (2) holds. Fix $u \in X$. Consider iteration $i \in [k]$. Note that set P_i satisfies property (b) of Theorem 3.4.1 regardless of what set X_i we have in the beginning of iteration i. That is, for every set $Y \subset X$ and $u \in Y$, we have

$$\sum_{v \in \mathrm{Ball}(u,R) \cap Y} \left(\Pr\left\{\delta_{P_i}(u,v) = 1 \,|\, X_i = Y\right\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_{P_i}(u,v) = 1 \,|\, X_i = Y\} \right)^+ \lesssim \beta^q \cdot \mathbb{E}\left[\sum_{v \in \mathrm{Ball}(u,2R) \cap Y} \frac{d(u,v)}{R} \cdot \vee_{P_i}(u,v) \,|\, X_i = Y \right]. \quad (3.7)$$

We observe that inequality (3.7) can be written as follows (for all $u \in X$).

$$\sum_{v \in \mathrm{Ball}(u,R)} \left(\Pr\left\{\delta_{P_i}(u,v) = 1 \text{ and } u,v \in X_i \,|\, X_i = Y\right\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_{P_i}(u,v) = 1 \text{ and } u,v \in X_i \,|\, X_i = Y\} \right)^+ \lesssim \beta^q \cdot \mathbb{E}\left[\sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_{P_i}(u,v) \cdot \mathbb{1}\left\{u,v \in X_i\right\} \,|\, X_i = Y \right]. \quad (3.8)$$

If $u \notin Y$, then all terms in (3.8) are equal to 0, and the inequality trivially holds. If $u \in Y$, then corresponding terms in (3.7) and (3.8) with $v \in Y$ are equal to each other; all terms in (3.8) with $v \notin Y$ are equal to 0. Denote the event that $u, v \in X_i$ by $\mathcal{E}_{vi}$ (that is, $\mathcal{E}_{vi}$ happens if both points u and v are not clustered at the beginning of iteration i). We take the expectation of (3.8) over $X_i = Y$ and add up the inequalities over all $i \in [k]$. Using the

subaddivity of function $x \mapsto x^+$, we obtain

$$\begin{aligned}
&\sum_{v \in \mathrm{Ball}(u,R)} \left(\sum_{i \in [k]} \Pr\left\{\delta_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\right\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\} \right)^+ \\
&\leq \sum_{\substack{v \in \mathrm{Ball}(u,R) \\ i \in [k]}} \left(\Pr\left\{\delta_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\right\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\} \right)^+ \\
&\lesssim \beta^q \cdot \mathbb{E}\left[\sum_{\substack{v \in \mathrm{Ball}(u,2R) \\ i \in [k]}} \frac{d(u,v)}{R} \cdot \vee_{P_i}(u,v) \cdot \mathbb{1}\{\mathcal{E}_{vi}\} \right].
\end{aligned} \tag{3.9}$$

Now consider any $v \in X \setminus \{u\}$. If u and v are separated by the partitioning $\mathcal{P}$, then they are separated at some iteration i. That is, for some $i \in [k]$:

- $\mathcal{E}_{vi}$ happens (in other words, u and v are not clustered at the beginning of iteration i)
- $\delta_{P_i}(u,v) = 1$ (exactly one of them gets clustered in iteration i)

Further, there is exactly one i such that both events above happen. On the other hand, if u and v are not separated by $\mathcal{P}$ then $\delta_{P_i}(u,v) = 0$ for all $i \in [k]$. We conclude that

$$\mathbb{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} = \sum_{i \in [k]} \mathbb{1}\left\{\delta_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\right\}. \tag{3.10}$$

In particular, the expectations of the expressions on both sides of (3.10) are equal:

$$\Pr\{\mathcal{P}(u) \neq \mathcal{P}(v)\} = \sum_{i \in [k]} \Pr\left\{\delta_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\right\}. \tag{3.11}$$

Now consider the first iteration i at which at least one of the vertices u and v gets clustered. Note that (i) event $\mathcal{E}_{vi}$ happens and (ii) $\vee_{P_i}(u,v) = 1$ (that is, (i) both points u and v are not clustered at the beginning of iteration i; (ii) but at least one of them gets clustered in iteration i). Further, for $j < i$, $\vee_{P_i}(u,v) = 0$ and for $j > i$, $\mathcal{E}_{vj}$ does not happen. We conclude

that event "$\vee_{P_i}(u,v) = 1$ and $\mathcal{E}_{vi}$" happens exactly for one value of $i \in [k]$. Therefore,

$$\sum_{i \in k} \vee_{P_i}(u,v) \cdot \mathbb{1}\{\mathcal{E}_{vi}\} = 1 \tag{3.12}$$

and

$$\sum_{i \in [k]} \Pr\{\vee_{P_i}(u,v) = 1 \text{ and } \mathcal{E}_{vi}\} = \sum_{i \in [k]} \mathbb{E}[\vee_{P_i}(u,v)\mathbb{1}\{\mathcal{E}_{vi}\}] = 1. \tag{3.13}$$

Plugging (3.11) and (3.13) into (3.9), we obtain

$$\sum_{v \in \text{Ball}(u,R)} \left(\Pr\left\{\mathcal{P}(u) = \mathcal{P}(v)\right\} - D_\beta \frac{d(u,v)}{R} \right)^+ \lesssim \beta^q \cdot \mathbb{E}\Big[\sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R} \Big].$$

We conclude that property (2) holds. Next, we show that property (3) holds for every $u \in X$. As in the analysis of property (2), we consider some iteration i. Then property (c) of Theorem 3.4.1 guarantees that if $u \in X_i$ then

$$\sum_{i \in [k]} \sum_{v \in \text{Ball}(u,r) \cap X_i} \delta_{P_i}(u,v) \lesssim \sum_{i \in [k]} \beta \cdot D_\beta^2 \cdot \left(\sum_{v \in \text{Ball}(u,2R) \cap X_i} \frac{d(u,v)}{R} \cdot \vee_{P_i}(u,v) \right) \tag{3.14}$$

We rewrite (3.14) as follows:

$$\sum_{i \in [k]} \sum_{v \in \text{Ball}(u,r)} \delta_{P_i}(u,v) \cdot \mathbb{1}\{\mathcal{E}_{vi}\} \lesssim \sum_{i \in [k]} \beta \cdot D_\beta^2 \cdot \left(\sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_{P_i}(u,v) \cdot \mathbb{1}\{\mathcal{E}_{vi}\} \right).$$

Note that this inequality holds for all $u \in X$: if $u \in X_i$, it is equivalent to (3.14); if $u \notin X_i$, then both sides are equal to 0, and the inequality trivially holds. Using formulas (3.10) and

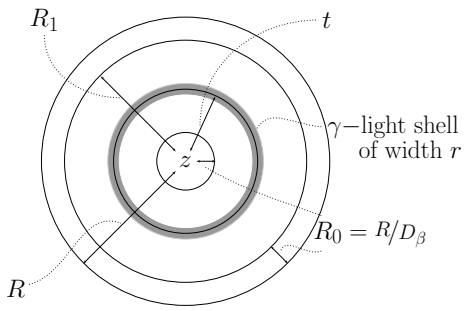

Figure 3.3: Light Ball
$R > r > 0$, $q \geq 1$, $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$, $R_0 = R/D_\beta$, $R_1 = R - R_0$.

(3.12), we get

$$\sum_{v \in \text{Ball}(u,r)} \mathbb{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} \lesssim \beta \cdot D_\beta^2 \cdot \sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R}.$$

Therefore, property (3) holds. □

3.6 Proof of Theorem 3.4.1

In Section 3.4.1, we describe an iterative approach to finding a probabilistic metric decomposition for Theorem 3.2.1. In this section, we show how to find one cluster P of the partitioning. Given a metric space (X, d) and positive numbers r and R, our algorithm selects a subset $P \subseteq X$ that satisfies the three properties listed in Theorem 3.4.1. Recall that $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$, $R_0 = R/D_\beta$, $R_1 = R - R_0$ and $\rho_q(\beta) = (1/\beta)^{q+1}$ (see Figure 3.3). In this proof, we assume that $\beta = r/R$ is sufficiently small (i.e, $\beta \leq \beta_q^*$ for some small $\beta_q^* = \Theta\big(\frac{1}{(q\ln(q+1)}\big)$ and, consequently, $R_0 = R/D_\beta$ is also small. Specifically, we assume that $r < R_0 < R_1 < R$ and $R_0 + r < R_1/100$.

Our algorithm for selecting the cluster P starts by picking a pivot point z that has the most points within a ball of small radius R_0. That is, z is the optimizer to the following

expression:

$$z = \arg\max_{u \in X} |\operatorname{Ball}(u, R_0)|. \tag{3.15}$$

The algorithm then checks if the ball of a larger radius, R_1, around z has significantly more points in it in comparison to the ball of radius R_0 around z. If the ratio of the number of points in these two balls exceeds $\rho_q(\beta)$, the algorithm selects the set of points $\operatorname{Ball}(z, R_1)$ as our cluster P. We refer to this case as the "Heavy Ball" case. In Section 3.6.2, we show that this set P satisfies the properties of Theorem 3.4.1.

If, however, $|\operatorname{Ball}(z, R_1)| < \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$, then the algorithm outputs cluster $P = \operatorname{Ball}(z, t)$ where $t \in (0, R]$ is chosen as follows. First, the algorithm finds the set S of all radii $s \in (3R_0, R_1]$ for which the set $P = \operatorname{Ball}(z, s)$ satisfies Definition 3. Then, it chooses a random radius t in S (non-uniformly) so that random set $P = \operatorname{Ball}(z, t)$ satisfies property (b) of Theorem 3.4.1. In Section 3.6.4, we discuss how to find the set S and show that $\mu(S) \geq R/2$ (where $\mu(S)$ is the Lebesgue measure of set S). Finally, in Section 3.6.5, we describe a procedure for choosing a random radius t in S.

3.6.1 Useful Observations

In this section, we prove several lemmas which we will use for analyzing both the "Heavy Ball" and "Light Ball" cases. First, we show an inequality that will help us lower bound the right hand sides in inequalities (b) and (c) of Theorem 3.4.1.

Lemma 3.6.1. *Assume that z is chosen according to (3.15). Consider t in $(3R_0, R_1]$ and u in X with $d(z, u) \in [2R_0, R]$. Let $P = \operatorname{Ball}(z, t)$. Denote:*

$$Y_P = \sum_{v \in \operatorname{Ball}(u, 2R)} \frac{d(u, v) \vee_P (u, v)}{R}.$$

Then, $|P| \leq 2D_\beta Y_P$.

Remark: Note that in Theorem 3.4.1, the right side of inequality (b) equals $\beta^q \mathbb{E}[Y_P]$, and the right side of inequality (c) equals $\beta \cdot D_\beta^2 \cdot Y_p$.

Proof. Observe that $P \subset \text{Ball}(u, 2R)$. Hence,

$$Y_P = \sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v) \vee_P (u,v)}{R} \geq \sum_{v \in P} \frac{d(u,v)}{R}.$$

For every $v \in P \setminus \text{Ball}(u, R_0)$, we have $\frac{d(u,v)}{R} \geq \frac{R_0}{R} = D_\beta^{-1}$. Thus,

$$Y_P \geq D_\beta^{-1} \, |P \setminus \text{Ball}(u, R_0)|.$$

We need to lower bound the size of $P \setminus \text{Ball}(u, R_0)$. On the one hand, we have

$$|P \setminus \text{Ball}(u, R_0)| \geq |P| - |\,\text{Ball}(u, R_0)| \geq |P| - |\,\text{Ball}(z, R_0)|.$$

Here, we used that $\text{Ball}(z, R_0)$ is the largest ball of radius R_0 in X. On the other hand, $\text{Ball}(z, R_0) \subset P \setminus \text{Ball}(u, R_0)$, since $d(z, u) \geq 2R_0$. Thus, $|P \setminus \text{Ball}(u, R_0)| \geq |\,\text{Ball}(z, R_0)|$. Combining two bounds on $|P \setminus \text{Ball}(u, R_0)|$, we get the desired inequality $|P \setminus \text{Ball}(u, R_0)| \geq |P|/2$. □

We now provide a lemma that will help us verify property (b) of Theorem 3.4.1 for that point u.

Lemma 3.6.2. *Consider an arbitrary probability distribution of* t *in* $(3R_0, R_1]$. *Let* $P = \text{Ball}(z, t)$, *where* z *is chosen according to (3.15). If for each point* $u \in \text{Ball}(z, R)$ *at least one of the following two conditions holds, then* P *satisfies property (b) of Theorem 3.4.1 for all points* u *in* X.

Condition I:

$$\Pr\{t \geq d(z,u) - R_0\} \lesssim \beta^{q+1} \cdot \frac{\mathbb{E}|\operatorname{Ball}(z,t)|}{|\operatorname{Ball}(z,R_0)|}. \tag{3.16}$$

Condition II: For every $v \in \operatorname{Ball}(u,R_0)$, we have

$$\Pr\left\{\delta_P(u,v)=1\right\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_P(u,v)=1\} \lesssim \beta^{q+1}. \tag{3.17}$$

Remark: This lemma makes the argument about the distribution of t from the proof overview section (Section 3.4) more precise. As we discuss in subsection Light Ball 3.4.2, we have chosen the distribution of t (Cumulative distribution function F, Definition 5) to satisfy two properties: (Property I) the probability that u and v are separated by P is upper bounded by the probability that u or v is in P times $O(D_\beta)$; and (Property II) The probability that t is close to $\pi_S^{inv}(R/2)$ is small. Thus, Condition I of Lemma 3.6.2 holds for u with $d(z,u)$ that are sufficiently close to $\pi_S^{inv}(R/2)$, and Condition II holds for u with for smaller values of $d(z,u)$.

Proof. Consider one term from the left hand side of property (b) of Theorem 3.4.1 for some u in X:

$$\left(\Pr\left\{\delta_P(u,v)=1\right\} - D_\beta \frac{d(u,v)}{R} \cdot \Pr\{\vee_P(u,v)=1\}\right)^+.$$

Note that $\{\delta_P(u,v)=1\}$ denotes the event that *exactly one of the points u and v* lies in P; whereas $\{\vee_P(u,v)=1\}$ denotes the event that *at least* one of u and v lies in P. Thus, $\Pr\{\vee_P(u,v)=1\} \geq \Pr\{\delta_P(u,v)=1\}$. Hence, this expression is positive only if $D_\beta \cdot d(u,v)/R < 1$, which is equivalent to

$$d(u,v) < R/D_\beta = R_0.$$

Thus, in the left hand side of property (b), we can consider only v in $\operatorname{Ball}(u,R_0)$ (rather than $\operatorname{Ball}(u,R)$). Moreover, if $d(z,u) > R$, then for all $v \in \operatorname{Ball}(u,R_0)$, we have $d(z,v) > R - R_0 =$

R_1 and, consequently, $\delta_P(u,v) = 0$. Therefore, for such u, the left hand side of property (b) equals 0, and the inequality (b) holds trivially. We will thus assume that $d(z,u) \le R$ (which is equivalent to $u \in \mathrm{Ball}(z,R)$). Similarly, since $t > 3R_0$, we will assume that $d(z,u) \ge 2R_0$ (otherwise, $u \in P$ and every $v \in \mathrm{Ball}(u,R_0)$ is in P, and thus $\delta_P(u,v) = 0$).

We now show that if Condition I or II of Lemma 3.6.2 holds for $u \in \mathrm{Ball}(z,R)$ then property (b) is satisfied for that u.

I. Suppose, the first condition holds for $u \in \mathrm{Ball}(z,R)$. If $\delta_P(u,v) = 1$ then either $u \in P,\ v \notin P$ or $v \in P,\ u \notin P$. In the former case, $t \ge d(z,u)$; in the latter case, $t \ge d(z,v) \ge d(z,u) - R_0$. In either case, $t \ge d(z,u) - R_0$. Using that $|\mathrm{Ball}(u,R_0)| \le |\mathrm{Ball}(z,R_0)|$ by our choice of z (see (3.15)), we bound the left hand side of (b) as follows

$$\begin{aligned}\sum_{v \in \mathrm{Ball}(u,R_0)} \Pr\{\delta_P(u,v) = 1\} &\le |\mathrm{Ball}(u,R_0)| \Pr\{t \ge d(z,u) - R_0\} \\ &\le |\mathrm{Ball}(z,R_0)| \Pr\{t \ge d(z,u) - R_0\}\end{aligned}$$

We now use the inequality from Condition I to get the bound

$$\sum_{v \in \mathrm{Ball}(u,R_0)} \Pr\{\delta_P(u,v) = 1\} \lesssim \beta^{q+1} \mathbb{E}|\mathrm{Ball}(z,t)| = \beta^{q+1}\mathbb{E}|P|.$$

Finally, by Lemma 3.6.1, we have the following bound on $\beta^{q+1}\mathbb{E}|P|$:

$$\beta^{q+1}\mathbb{E}|P| \le \beta^{q+1} \cdot 2D_\beta \mathbb{E}[Y_P] \le \beta^q \mathbb{E}[Y_P]. \tag{3.18}$$

Here, we used that $2\beta D_\beta = 2r/R_0 < 1$ by our choice of β_q^*. The right hand side of the inequality in property (b) equals $\beta^q \mathbb{E}[Y_P]$. Thus, property (b) holds.

II. Suppose now that the second condition holds for $u \in \mathrm{Ball}(z,R)$. Then, each term in the left hand side of (b) is upper bounded by $O(\beta^{q+1})$. Hence, the entire sum is upper bounded

by $O(\beta^{q+1}|\operatorname{Ball}(u, R_0)|)$, which in turn is upper bounded by $O(\beta^{q+1}|\operatorname{Ball}(z, R_0)|)$. Then,

$$\beta^{q+1}|\operatorname{Ball}(z, R_0)| \leq \beta^{q+1}|\operatorname{Ball}(z, t)| = \beta^{q+1}\mathbb{E}|P| \leq \beta^q \mathbb{E}[Y_P].$$

The last inequality follows from (3.18). We conclude that property (b) of Theorem 3.4.1 holds. □

3.6.2 Heavy Ball Case

In this subsection, we analyze the case when $|\operatorname{Ball}(z, R_1)| \geq \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$. If this condition is met, then the algorithm outputs $P = \operatorname{Ball}(z, R_1)$. We will show that Theorem 3.4.1 holds for such a cluster P.

We first prove properties (a) and (b). Since the radius of P is $R_1 \leq R$, its diameter is at most $2R$. So property (a) of Theorem 3.4.1 holds. To show property (b), we apply Lemma 3.6.2 (item I) with $t = R_1$. Trivially, $\Pr\{t \geq d(z, u) - R_0\} \leq 1$ and $\mathbb{E}|\operatorname{Ball}(z, t)| = |\operatorname{Ball}(z, R_1)| \geq \rho_q(\beta)|\operatorname{Ball}(z, R_0)|$. Thus, (3.6.2) is satisfied and property (b) also holds.

We now show property (c) of Theorem 3.4.1. Observe that if $d(z, u) \notin [R_1 - r, R_1 + r]$, then $\delta_P(u, v) = 0$ for all $v \in \operatorname{Ball}(u, r)$ (because $P = \operatorname{Ball}(z, R_1)$). Hence, for such u, property (c) holds. Thus, we assume that $u \in [R_1 - r, R_1 + r] \subseteq [2R_0, R]$.

We bound the left hand side of (c) as follows:

$$\sum_{v \in \operatorname{Ball}(u,r)} \delta(u, v) \leq |\operatorname{Ball}(u, r)| \leq |\operatorname{Ball}(u, R_0)| \leq |\operatorname{Ball}(z, R_0)|,$$

here we first use that $r \leq R_0$ and then that z satisfies (3.15). Since we are in the Heavy Ball Case, we have $|P| \geq \rho_q(\beta)|\operatorname{Ball}(z, R_0)|$. Therefore,

$$\sum_{v \in \operatorname{Ball}(u,r)} \delta(u, v) \leq |P|/\rho_q(\beta).$$

By Lemma 3.6.1, the right hand side is upper bounded by

$$2D_\beta Y_P/\rho_q(\beta) = 2D_\beta\, \beta^{q+1}\, Y_P \lesssim \beta D_\beta^2\, Y_P.$$

The right hand side of (c) equals $\beta D_\beta^2\, Y_P$. Hence, property (c) is satisfied.

Thus we have shown that Theorem 3.4.1 holds for the case of Heavy Balls. To complete the proof, we show that Theorem 3.4.1 also holds for the case of Light Balls – we give this proof in Section 3.6.3.

3.6.3 Light Ball Case

We now consider the case when $|\operatorname{Ball}(z, R_1)| \leq \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$. Recall that S is the set of all radii $s \in (3R_0, R_1]$ for which property (c) of Theorem 3.4.1 holds (Definition 3). The set S can be found in polynomial time since the number of distinct balls $\operatorname{Ball}(z, s)$ is upper bounded by the number of points in the metric space. We now recall map π_S used in Algorithm 5.

Map π_S. In Section 3.4.2, we define a measure preserving transformation π_S that maps a given measurable set $S \subset [0, R]$ to the interval $[0, \mu(S)]$ (Definition 4). We need this transformation in Algorithm 5. If S is the union of several disjoint intervals (as in our algorithm) then π_S simply pushes all intervals to the left so that every two consecutive intervals touch each other. We show the following lemma.

Lemma 3.6.3. *For any measurable set S, π_S is a continuous non-decreasing 1-Lipschitz function, and π_S^{inv} is a strictly increasing function defined for all y in $[0, \mu(S)]$. Moreover, there exists a set Z_0 of measure zero such that for all $y \in [0, \mu(S)] \setminus Z_0$, we have $\pi_S^{inv}(y) \in S$.*

Proof. Note that $\pi_S^{inv}(y)$ is a right inverse for $\pi_S(x)$: $\pi_S(\pi_S^{inv}(y)) = y$ (but not necessarily a

left inverse). Let

$$I_S(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{otherwise} \end{cases}$$

be the indicator function of set S. Then $\pi_S(x) = \int_0^x I_S(t)dt$ (we use Lebesgue integration here). Since $0 \le I_S(t) \le 1$, function π_S is non-decreasing, 1-Lipschitz, and absolutely continuous. By the Lebesgue differentiation theorem, $\pi_S(x)$ is almost everywhere differentiable and $\frac{d\pi_S(x)}{dx} = I_S(x)$ almost everywhere. Let $X_0 = [0, R] \setminus S$ and $Z_0 = \pi_S(X_0)$. Since π_S is absolutely continuous and $I_S(x) = 0$ for $x \in X_0$, we have

$$\mu(Z_0) \le \int_{X_0} \frac{d\pi_S(x)}{dx} dx = \int_{X_0} I_S(x)dx = 0.$$

Now if $y \notin Z_0$, then $\pi_S(\pi_S^{inv}(y)) = y \notin Z_0$, thus $\pi_S^{inv}(y) \notin X_0$ or, equivalently, $\pi_S^{inv}(y) \in S$, as required.

Finally, we verify that π_S^{inv} is strictly increasing. Consider $a, b \in [0, \mu(S)]$ with $a < b$. Note that $a = \pi_S(\pi_S^{inv}(a))$ and $b = \pi_S(\pi_S^{inv}(b))$. Thus, $\pi_S(\pi_S^{inv}(a)) < \pi_S(\pi_S^{inv}(b))$. Since π_S is non-decreasing, $\pi_S^{inv}(a) < \pi_S^{inv}(b)$. □

Note that if S is a union of finitely many disjoint open intervals, then Z_0 is the image of the endpoints of those intervals under π_S.

3.6.4 Clusters Satisfying Property (c) of Theorem 3.4.1

We first show that if $|\operatorname{Ball}(z, R_1)| < \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$, then $\mu(S) \ge R/2$. To this end, we define a ball with a γ-light shell of width r.

Definition 6. *We say that the ball of radius $t \ge r$ around z has a γ-light shell of width r if*

$$|\operatorname{Ball}(z, t + r)| - |\operatorname{Ball}(z, t - r)| \le \gamma \int_0^{t-r} |\operatorname{Ball}(z, x)|\, dx.$$

We let S_γ be the set of all radii t in the range $(3R_0, R_1]$ such that $\text{Ball}(z,t)$ has a γ-light shell of width r. We now show that (a) $S_\gamma \subset S$ and (b) $\mu(S_\gamma) \geq R/2$ for $\gamma = 25r/R_0^2$. and, therefore, $\mu(S) \geq R/2$.

Lemma 3.6.4. *We have $S_\gamma \subset S$.*

Proof. Consider a number t from S_γ and the ball of radius t around z: $P = \text{Ball}(z,t)$. Let us pick an arbitrary point u. We are going to prove that inequality (3.5) holds and therefore $t \in S$. Consider $v \in \text{Ball}(u,r)$. Observe that $\delta_P(u,v) = 1$ only if both u and v belong to the r neighborhoods of P and $X \setminus P$. Thus, if $\delta_P(u,v) = 1$, we must have $d(z,u), d(z,v) \in [t-r, t+r]$. If $d(z,u) \notin [t-r, t+r]$, then the left side of (3.5) equals 0, and we are done. Hence, we can assume that $d(z,u) \in [t-r, t+r]$.

Using the observation above, we bound the left hand side of (3.5) as

$$\sum_{v \in \text{Ball}(u,r)} \delta_P(u,v) \leq |\,\text{Ball}(z, t+r)| - |\,\text{Ball}(z, t-r)|.$$

We now need to lower bound the right hand side of (3.5). Note that $\text{Ball}(u, 2R)$ contains $\text{Ball}(z,t)$, since

$$d(z,u) \leq t + r \leq R_1 + r = R - R_0 + r < R,$$

and $t < R$. Thus,

$$\sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R} \vee_p (u,v) \geq \frac{1}{R} \sum_{v \in \text{Ball}(z,t)} d(u,v) \vee_p (u,v).$$

For all $v \in \text{Ball}(z,t) \equiv P$, we have $\vee_P(u,v) = 1$. Hence,

$$\sum_{v \in \text{Ball}(z,t)} d(u,v) \vee_p (u,v) = \sum_{v \in \text{Ball}(z,t)} d(u,v) \tag{3.19}$$

By the triangle inequality, we have

$$d(u,v) \geq (d(z,u) - d(z,v))^+ \geq ((t-r) - d(z,v))^+.$$

Observe that

$$((t-r) - d(z,v))^+ = \int_0^{t-r} \mathbb{1}\big\{d(z,v) \leq x\big\}\, dx.$$

Hence, (3.19) is lower bounded by

$$\sum_{v \in \mathrm{Ball}(z,t)} \int_0^{t-r} \mathbb{1}\big\{d(z,v) \leq x\big\}\, dx = \int_0^{t-r} \sum_{v \in \mathrm{Ball}(z,t)} \mathbb{1}\big\{d(z,v) \leq x\big\}\, dx = \int_0^{t-r} |\,\mathrm{Ball}(z,x)|\, dx.$$

Since the ball of radius t has a γ-light shell of width r, the expression above is, in turn, lower bounded by

$$\frac{|\,\mathrm{Ball}(z,t+r)| - |\,\mathrm{Ball}(z,t-r)|}{\gamma}.$$

Thus, the right hand side of inequality (3.5) is lower bounded by

$$\frac{25\beta D_\beta^2}{R} \cdot \frac{|\,\mathrm{Ball}(z,t+r)| - |\,\mathrm{Ball}(z,t-r)|}{\gamma}.$$

This completes the proof of Lemma 3.6.4, since

$$\frac{25\beta D_\beta^2}{R} \cdot \frac{1}{\gamma} = \frac{25\beta D_\beta^2 R_0^2}{25R \cdot r} = \frac{(r/R)\, D_\beta^2 \left(R/D_\beta\right)^2}{Rr} = 1.$$

□

Lemma 3.6.5. *We have* $\mu\big(S_\gamma\big) \geq R/2$.

To prove this lemma, we use the following result from Appendix 4.2.

Lemma 3.6.6. *Consider a non-decreasing function* $\Phi : [0,R] \to \mathbb{R}$ *with* $\Phi(0) = 1$ *and* $R > 0$. *Let* $r \in (0,R]$ *and* $\gamma \leq (0, 1/r]$. *Then, for the subset* S *of numbers* $t \in [0, R-r]$ *for which*

inequality

$$\Phi(t+r) \geq \Phi(t) + \gamma \int_0^t \Phi(x)dx \tag{3.20}$$

holds, we have $\Phi(R) \geq e^{\eta\mu(S)-1}$*, where* $\eta = \sqrt{\gamma/(e-1)r}$*, and* $\mu(S)$ *is the measure of set* S*.*

Proof of Lemma 3.6.5. We apply Lemma 3.6.6 to the function

$$\Phi(t) = \frac{|\operatorname{Ball}(z, t+3R_0)|}{|\operatorname{Ball}(z, 3R_0)|}$$

with parameters $r' = 2r$, $R' = R_1 - 3R_0 - r$, and $\gamma = 25r/R_0^2$. Note that to be able to apply Lemma 3.6.6 we need $\gamma < 1/r'$ which is equivalent to $\beta D_\beta < 1/5\sqrt{2}$. The latter holds due to β being sufficiently small, i.e., $\beta \leq \Theta\big(\frac{1}{(q\ln(q+1)}\big)$. Observe that $\Phi(0) = 1$ and

$$\Phi(R') \leq \frac{|\operatorname{Ball}(z, R_1)|}{|\operatorname{Ball}(z, 3R_0)|} \leq \frac{\rho_q(\beta)|\operatorname{Ball}(z, R_0)|}{|\operatorname{Ball}(z, R_0)|} = \rho_q(\beta).$$

Here, we used that the $\operatorname{Ball}(z, R_1)$ is light. From Lemma 3.6.6, we get that $\Phi(R') \geq e^{\eta'\mu(S')-1}$, where $\eta' = \sqrt{\gamma/(e-1)r'}$, and S' is the set of t for which Inequality (3.20) holds. Thus,

$$\begin{aligned}
\mu(S') &\leq \frac{1+\ln\Phi(R')}{\eta'} \leq \frac{1+\ln\rho_q(\beta)}{\eta'} = \frac{1+D_\beta/2}{\eta'} \\
&= \sqrt{\frac{(e-1)r'}{\gamma}} \cdot (1+D_\beta/2) \\
&= \sqrt{\frac{2(e-1)r}{25r}} \cdot R_0 \cdot (1+D_\beta/2) \\
&= \sqrt{\frac{2(e-1)}{25}} \cdot (R_0 + R/2) < 0.4(R+R_0).
\end{aligned}$$

where we used $R_0 \cdot D_\beta = R$ and that $\sqrt{2(e-1)} < 2$. Therefore for the measure of the set $S'' = [0, R'] \setminus S'$ is at least $\mu(S'') \geq ((R - R_0) - 3R_0 - r) - 0.4(R + R_0) \geq R/2$. Here, we relied on our assumption that $R_0 + r < R_1/100$.

We claim that $S'' + 3R_0 + r \subset S_\gamma$. Consider an arbitrary $t \in S''$. First, observe that

$t + 3R_0 + r \in (3R_0, R_1]$. Then,

$$\frac{|\operatorname{Ball}(z, t + 3R_0 + r')|}{|\operatorname{Ball}(z, 3R_0)|} - \frac{|\operatorname{Ball}(z, t + 3R_0)|}{|\operatorname{Ball}(z, 3R_0)|} = \Phi(t + r') - \Phi(t)$$
$$< \gamma \int_0^t \Phi(x)dx = \gamma \int_0^t \frac{|\operatorname{Ball}(z, x + 3R_0)|}{|\operatorname{Ball}(z, 3R_0)|}dx.$$

For $t' = t + 3R_0 + r$, we get

$$|\operatorname{Ball}(z, t' + r)| - |\operatorname{Ball}(z, t' - r)| < \gamma \int_0^{t'-3R_0-r} |\operatorname{Ball}(z, x + 3R_0)|dx$$
$$= \gamma \int_{3R_0}^{t'-r} |\operatorname{Ball}(z, x)|dx < \gamma \int_0^{t'-r} |\operatorname{Ball}(z, x)|dx.$$

Thus, $t' \in S_\gamma$. This finishes the proof. □

Lemma 3.6.5 together with Lemma 3.6.4 imply the following corollary.

Corollary 3.6.7. *Let S be the set defined in Definition 3. Then, $\mu(S) \geq R/2$.*

3.6.5 Clusters Satisfying Property (b) of Theorem 3.4.1

We now show how to choose a random $t \in S$, so that the random cluster $P = \operatorname{Ball}(z, t)$ satisfies property (b) of Theorem 3.4.1. We first choose a random $x \in [0, R/2]$ with the cumulative distribution function $F(x)$ defined in Definition 5, and then let $t = \pi_S^{inv}(x)$, where $S \subset (3R_0, R_1]$ is the set obtained in the previous section. Note that by Lemma 3.6.3, $t = \pi_S^{inv}(x) \in S$ with probability 1, since $\Pr\{x \in Z_0\} = 0$ (see Lemma 3.6.3).

To show that property (b) is satisfied, we verify that for every u in $\operatorname{Ball}(z, R)$, Condition I or Condition II of Lemma 3.6.2 holds.

Pick a point u in $\operatorname{Ball}(z, R)$. We consider two cases: $\pi_S(d(z, u)) > R/2 - R_0$ and $\pi_S(d(z, u)) \leq R/2 - R_0$. We prove that u satisfies Condition I of Lemma 3.6.2 in the former

case and Condition II in the latter case.

First case: $\pi_S(d(z,u)) > R/2 - R_0$. Write,

$$\Pr\{t \geq d(z,u) - R_0\} = \Pr\{x \geq \pi_S(d(z,u) - R_0)\}.$$

Since π_S is a 1-Lipschitz function, we have

$$\pi_S(d(z,u) - R_0) \geq \pi_S(d(z,u)) - R_0 \geq R/2 - 2R_0.$$

Therefore,

$$\Pr\{t \geq d(z,u) - R_0\} \leq 1 - F(R/2 - 2R_0).$$

We prove the following claim.

Claim 3.6.8. *We have*

$$1 - F(R/2 - 2R_0) \lesssim \beta^{q+1}.$$

Proof. Write:

$$F(R/2 - 2R_0) = \frac{1 - e^{\frac{-R}{2R_0}} \cdot e^{\frac{2R_0}{R_0}}}{1 - e^{\frac{-R}{2R_0}}} = \frac{1 - e^2 e^{-D\beta/2}}{1 - e^{-D\beta/2}}.$$

Note that $e^{-D\beta/2} = \beta^{q+1}$. Then,

$$1 - F(R/2 - 2R_0) = \frac{(e^2 - 1)}{1 - \beta^{q+1}} \cdot \beta^{q+1}.$$

Since the denominator of the right hand side is greater than 1/2 (recall that we assume that β is sufficiently small), we have $1 - F(R/2 - 2R_0) \lesssim \beta^{q+1}$. □

Claim 3.6.8 finishes the analysis of the first case, since $|\operatorname{Ball}(z,t)|/|\operatorname{Ball}(z,R_0)| \geq 1$ for every value of $t \geq R_0$.

Second case: $\pi_S(d(z,u)) \leq R/2 - R_0$. In this case, for every $v \in \text{Ball}(u, R_0)$, we have

$$\pi_S(d(z,v)) \leq \pi_S(d(z,u) + R_0) \leq R/2.$$

Here, we used that π_S is a 1-Lipschitz function. We claim that inequality (3.17) holds for every two points $v_1, v_2 \in X$ with $\pi_S(d(z,v_1)), \pi_S(d(z,v_2)) \leq R/2$ and $d(v_1,v_2) \leq R_0$. In particular, it holds for $v_1 = u$ and $v_2 = v$. Without loss of generality assume, that $d(z,v_1) \leq d(z,v_2)$. Then,

$$\begin{aligned}
\Pr\{\delta_P(v_1,v_2) = 1\} &= \Pr\{d(z,v_1) \leq t < d(z,v_2)\} \\
&= \Pr\{\pi_S(d(z,v_1)) \leq x < \pi_S(d(z,v_2))\} \\
&= F(\pi_S(d(z,v_2))) - F(\pi_S(d(z,v_1))).
\end{aligned}$$

Here, we used that random variable x has distribution function F. We show the following claim.

Claim 3.6.9. *For all $x_1 \leq x_2$ in the range $[0, R/2]$, we have*

$$F(x_2) - F(x_1) \leq D_\beta \cdot \frac{(x_2 - x_1)}{R} \cdot \big(1 - F(x_1) + 2\beta^{q+1}\big).$$

Proof. We have

$$\begin{aligned}
F(x_2) - F(x_1) &= \int_{x_1}^{x_2} F'(x)dx \\
&\leq (x_2 - x_1) \max_{x \in [x_1, x_2]} F'(x) \\
&= (x_2 - x_1) \cdot \frac{e^{-x_1/R_0}/R_0}{1 - e^{-R/2R_0}} \\
&= D_\beta \cdot \frac{(x_2 - x_1)}{R} \cdot \frac{e^{-x_1/R_0}}{1 - e^{-R/2R_0}}.
\end{aligned}$$

Here, we used that $R_0 = R/D_\beta$. We now need to upper bound the third term on the right hand side:

$$\frac{e^{-x_1/R_0}}{1-e^{-R/2R_0}} = 1 - \frac{(1-e^{-R/2R_0}) - e^{-x_1/R_0}}{1-e^{-R/2R_0}} = 1 - F(x_1) + \frac{e^{-R/2R_0}}{1-e^{-R/2R_0}}.$$

As in Claim 3.6.8, let us use that $e^{-R/2R_0} = \beta^{q+1}$ and $1 - \beta^{q+1} \geq 1/2$ to get

$$\frac{e^{-x_1/R_0}}{1-e^{-R/2R_0}} \leq 1 - F(x_1) + 2\beta^{q+1}.$$

Combining the bounds above, we get the following inequality:

$$F(x_2) - F(x_1) \leq D_\beta \cdot \frac{x_2 - x_1}{R} \cdot \Big(1 - F(x_1) + 2\beta^{q+1}\Big).$$

□

Using Claim 3.6.9 and the inequality

$$\pi_S(d(z,v_2))) - \pi_S(d(z,v_1) \leq d(z,v_2) - d(z,v_1) \leq d(v_1,v_2),$$

we derive the following upper bound

$$\Pr\{\delta_P(v_1,v_2) = 1\} \leq D_\beta \frac{d(v_1,v_2)}{R} \cdot \big(1 - F(\pi_S(d(z,v_1)))) + 2\beta^{q+1}\big).$$

Then,

$$\Pr\{\vee_P(v_1,v_2) = 1\} = \Pr\{d(z,v_1) \leq t\} = 1 - F(\pi_S(d(z,v_1))).$$

Therefore,

$$\Pr\{\delta_P(v_1, v_2) = 1\} - D_\beta \frac{d(v_1, v_2)}{R} \cdot \Pr\{\vee_P(v_1, v_2) = 1\} \leq 2D_\beta \frac{d(v_1, v_2)}{R} \beta^{q+1}.$$

Thus, the left hand side of (3.17) is upper bounded by

$$2D_\beta \cdot \beta^{q+1} \cdot \frac{d(v_1, v_2)}{R} \leq 2\beta^{q+1}.$$

Here, we use that $d(v_1, v_2) \leq R_0$ and $R_0 = R/D_\beta$.

3.7 Integrality Gap

In this section, we present an integrality gap example for the convex program (P) in Figure 3.1.

Proof of Theorem 1.4.8. Let $n = 1 + \lceil \sqrt{1/\alpha} \rceil$. Consider a complete graph on n vertices. Let P be a path of length $n - 1$. Denote its endpoints by s and t and the set of its edges by E_P. All edges in P are positive edges of weight 1. Edge (s, t) is a negative edge of weight 1. All other edges are positive edges of weight α.

The cost of the integral solution Clearly, every integral solution $\mathcal{P}$ should violate some edge $(u, v) \in P \cup \{(s, t)\}$ (since all these edges cannot be satisfied simultaneously). Thus, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-) \geq 1$ and $\|\mathrm{dis}(\mathcal{P}, E^+, E^-)\|_p \geq 1$.

The cost of the CP solution. We define the CP solution as follows. Denote the distance between u and v along P by $\mathrm{dist}_P(u, v)$. Let $x_{uv} = \mathrm{dist}_P(u, v)/(n - 1)$. Note that $x_{st} = 1$. The values of variables y_u are determined by constraints (P1) of the convex program.

Now we upper bound the contribution of every edge (u, v) (incident on u) to y_u in formula (P1). The contribution of $(u, v) \in E_P$ is $w_{uv}x_{uv} = 1 \cdot 1/(n - 1)$; the contribution of edge (s, t) is $w_{st}(1 - x_{st}) = 0$ (whether or not it is incident on u), and the contribution of every

other edge (u, v) is $w_{uv}x_{uv} \leq \alpha$. Since every vertex u is incident on at most 2 edges from E_P, we have $y_u \leq 2/(n-1) + \alpha n \lesssim \sqrt{\alpha}$. Now,

$$\|y\|_p \leq n^{1/p} \cdot \max_u |y_u| \lesssim n^{1/p}\alpha^{1/2} \lesssim \alpha^{1/2-1/(2p)}.$$

Integrality gap We conclude that the integrality gap is at least $\Omega((1/\alpha)^{1/2-1/(2p)})$. □

CHAPTER 4

DEFERRED PROOFS

4.1 Proof of Theorem 2.3.1

For the sake of completeness we include the proof of Theorem 2.3.1 (see [3] and [19]).

Proof of Theorem 2.3.1. Our first task is to express the cost of violations made by Algorithm 1 and the LP weight in terms of $ALG^{\sigma}(\cdot)$ and $LP^{\sigma}(\cdot)$, respectively. In order to do this, we consider the cost of violations made by the algorithm at each step.

Consider step t of the algorithm. Let V_t denote the set of active (yet unclustered) vertices at the start of step t. Let $w \in V_t$ denote the pivot chosen at step t. The algorithm chooses a set $S_t \subseteq V_t$ as a cluster and removes it from the graph. Notice that for each $u \in S_t$, the constraint imposed by each edge of type $(u, v) \in E^+ \cup E^-$ is satisfied or violated right after step t. Specifically, if (u, v) is a positive edge, then the constraint (u, v) is violated if exactly one of the vertices u, v is in S_t. If (u, v) is a negative constraint, then it is violated if both u, v are in S_t. Denote the weight of violated constraints at step t by ALG_t. Thus,

$$ALG_t = \sum_{\substack{(u,v)\in E^+ \\ u,v\in V_t}} \boldsymbol{w}_{uv} \cdot \mathbb{1}\left(u \in S_t, v \notin S_t \text{ or } u \notin S_t, v \in S_t\right) + \sum_{\substack{(u,v)\in E^- \\ u,v\in V_t}} \boldsymbol{w}_{uv} \cdot \mathbb{1}\left(u \in S_t, v \in S_t\right).$$

Similarly, we can quantify the LP weight removed by the algorithm at step t, which we denote by LP_t. We count the contribution of all edges $(u, v) \in E^+ \cup E^-$ such that $u \in S_t$ or $v \in S_t$. Thus,

$$LP_t = \sum_{\substack{(u,v)\in E^+ \\ u,v\in V_t}} \boldsymbol{w}_{uv} x_{uv} \cdot \mathbb{1}(u \in S_t \text{ or } v \in S_t) + \sum_{\substack{(u,v)\in E^- \\ u,v\in V_t}} \boldsymbol{w}_{uv}(1 - x_{uv}) \cdot \mathbb{1}(u \in S_t \text{ or } v \in S_t)$$

Note that the cost of the solution produced by the algorithm is the sum of the violations

across all steps, that is $ALG = \sum_t ALG_t$. Moreover, as every edge is removed exactly once from the graph, we can see that $LP = \sum_t LP_t$. We will charge the cost of the violations of the algorithm at step t, ALG_t, to the LP weight removed at step t, LP_t. Hence, if we show that $\mathbb{E}[ALG_t] \leq \rho \mathbb{E}[LP_t]$ for every step t, then we can conclude that the approximation factor of the algorithm is at most ρ, since

$$\mathbb{E}[ALG] = \mathbb{E}\left[\sum_t ALG_t\right] \leq \rho \cdot \mathbb{E}\left[\sum_t LP_t\right] = \rho \cdot LP.$$

We now express ALG_t and LP_t in terms of $cost(\cdot)$ and $lp(\cdot)$ which are defined in Section 2.3.1. This will allow us to group together the terms for each triplet u, v, w in the set of active vertices and thus write ALG_t and LP_t in terms of $ALG^\sigma(\cdot)$ and $LP^\sigma(\cdot)$, respectively.

For analysis, we assume that for each vertex $u \in V$, there is a positive (similar) self-loop, and thus we can define $cost(u, u \,|\, w)$ and $lp(u, u \,|\, w)$ formally as follows: $cost(u, u \,|\, w) = \Pr(u \in S, u \notin S \,|\, p = w) = 0$ and $lp(u, u \,|\, w) = x_{uu} \cdot \Pr(u \in S \,|\, p = w) = 0$ (recall that $x_{uu} = 0$).

$$\mathbb{E}[ALG_t \,|\, V_t] = \sum_{\substack{(u,v)\in E \\ u,v\in V_t}} \left(\frac{1}{|V_t|} \sum_{w \in V_t} \boldsymbol{w}_{uv} \cdot cost(u, v \,|\, w)\right) = \frac{1}{2|V_t|} \sum_{\substack{u,v,w\in V_t \\ u\neq v}} \boldsymbol{w}_{uv} \cdot cost(u, v \,|\, w) \tag{4.1}$$

$$\mathbb{E}[LP_t \,|\, V_t] = \sum_{\substack{(u,v)\in E \\ u,v\in V_t}} \left(\frac{1}{|V_t|} \sum_{w \in V_t} \boldsymbol{w}_{uv} \cdot lp(u, v \,|\, w)\right) = \frac{1}{2|V_t|} \sum_{\substack{u,v,w\in V_t \\ u\neq v}} \boldsymbol{w}_{uv} \cdot lp(u, v \,|\, w) \tag{4.2}$$

We divide the expressions on the right hand side by 2 because the terms $cost(u, v \,|\, w)$ and $lp(u, v \,|\, w)$ are counted twice. Now adding the contribution of terms $cost(u, u \,|\, w)$ and $lp(u, u \,|\, w)$ (both equal to 0) to (4.1) and (4.2), respectively and grouping the terms containing

u, v and w together, we get,

$$\begin{aligned}\mathbb{E}[ALG_t \mid V_t] =& \frac{1}{6|V_t|} \sum_{u,v,w \in V_t} \Big(\boldsymbol{w}_{uv} \cdot cost(u, v \mid w) + \boldsymbol{w}_{uw} \cdot cost(u, w \mid v) + \boldsymbol{w}_{wv} \cdot cost(w, v \mid u) \Big) \\ =& \frac{1}{6|V_t|} \sum_{u,v,w \in V_t} ALG^{\sigma}(x, y, z)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[LP_t \mid V_t] =& \frac{1}{6|V_t|} \sum_{u,v,w \in V_t} \Big(\boldsymbol{w}_{uv} \cdot lp(u, v \mid w) + \boldsymbol{w}_{uw} \cdot lp(u, w \mid v) + \boldsymbol{w}_{wv} \cdot lp(w, v \mid u) \Big) \\ =& \frac{1}{6|V_t|} \sum_{u,v,w \in V_t} LP^{\sigma}(x, y, z)\end{aligned}$$

Thus, if $ALG^{\sigma}(x, y, z) \leq \rho LP^{\sigma}(x, y, z)$ for all signatures and edge lengths x, y, z satisfying the triangle inequality, then $\mathbb{E}[ALG_t \mid V_t] \leq \rho \cdot \mathbb{E}[LP_t \mid V_t]$, and, hence, $\mathbb{E}[ALG] \leq \rho \cdot \mathbb{E}[LP]$ which finishes the proof. □

4.2 Proof of Lemma 3.6.6

We first prove Lemma 3.6.6 for the case when S is a measure zero set. Specifically, we show the following lemma.

Lemma 4.2.1. *Suppose, a non-decreasing function $\Phi : [0, R] \to \mathbb{R}$ with $\Phi(0) = 1$ satisfies the following inequality for all $t \in [0, R - r] \setminus Y_0$, where set Y_0 has measure zero:*

$$\Phi(t + r) \geq \Phi(t) + \gamma \int_0^t \Phi(x) dx, \tag{4.3}$$

for some $R > 0$, $r \in (0, R/2]$ and $\gamma \in (0, 1/r]$, then $\Phi(t) \geq \max\{e^{\eta t - 1}, 1\}$ for all $t \in [0, R]$, where $\eta = \sqrt{\gamma/(e-1)r}$. Consequently, we have $\Phi(R) \geq e^{\eta R - 1}$.

Proof. Since $\Phi(0) = 1$ and $\Phi(t)$ is non-decreasing, we have $\Phi(t) \geq 1$ for all $t \geq 0$. We now

prove that $\Phi(t) \geq e^{\eta t - 1}$. We establish this inequality by induction. The inductive hypothesis is that this inequality holds for $t \in [0, 1/\eta + ir] \cap [0, R]$ for integer $i \geq 0$. For $t \leq 1/\eta$, we have $\Phi(t) \geq 1 > e^{\eta t - 1}$. Thus, the inductive hypothesis holds for $i = 0$. Suppose, it holds for i, we prove it for $i + 1$.

First, consider an arbitrary $t^* \in [1/\eta, 1/\eta + (i+1)r] \cap [0, R] \setminus (Y_0 + r)$, where $Y_0 + r$ is the set Y_0 shifted right by r. Let $t = t^* - r$. Note that $t > 0$, since $r < 1/\eta$. Also, $t \notin Y_0$. Then, by the inductive hypothesis, we have $\Phi(x) \geq e^{\eta x - 1}$ for all $x \in [1/\eta, t]$. Using Inequality (4.3), we obtain the following bound

$$\begin{aligned}
\Phi(t^*) &= \Phi(t + r) \\
&\geq \Phi(t) + \gamma \int_0^{1/\eta} \Phi(x) dx + \gamma \int_{1/\eta}^{t} \Phi(x) dx \\
&\geq e^{\eta t - 1} + \gamma \int_0^{1/\eta} 1 \, dx + \gamma \int_{1/\eta}^{t} e^{\eta x - 1} \, dx \\
&= e^{\eta t - 1} + \gamma/\eta + \gamma/\eta \cdot (e^{\eta t - 1} - 1) \\
&= e^{\eta t - 1}(1 + \gamma/\eta).
\end{aligned}$$

Since $\eta = \sqrt{\gamma/(e-1)r}$, we have $\gamma/\eta = (e-1)\eta r$. Now, using the inequality $e^x \leq 1 + (e-1)x$ for $x \in [0, 1]$, we get

$$\Phi(t^*) \geq e^{\eta t - 1}(1 + \gamma/\eta) = e^{\eta t - 1}(1 + (e-1)\eta\delta) \geq e^{\eta t - 1} \cdot e^{\eta r} = e^{\eta(t+r) - 1} = e^{\eta t^* - 1}.$$

To finish the proof, we need to show that $\Phi(t^{**}) \geq e^{\eta t^{**} - 1}$ for $t^{**} \in [1/\eta, 1/\eta + (i+1)r] \cap [0, R] \cap (Y_0 + r)$. Since $Y_0 + r$ has measure zero, there exists an increasing sequence t_k^* of numbers in $[0, 1/\eta + (i+1)r] \cap ([0, R] \setminus (Y_0 + r))$ that tends to t^{**} as $k \to \infty$. Using that Φ is a non-decreasing function and $e^{\eta t - 1}$ is a continuous function, we have

$$\Phi(t^{**}) \geq \lim_{k \to \infty} \Phi(t_k^*) \geq \lim_{k \to \infty} e^{\eta t_k^* - 1} = e^{\eta t^{**} - 1}.$$

□

We now show that Lemma 4.2.1 implies Lemma 3.6.6. Loosely speaking, in the proof, we shift all intervals from the set S to the left to obtain a single interval $[0, \mu(S)]$. We then apply Lemma 4.2.1 to the transformed function.

Proof of Lemma 3.6.6. Let π_S and π_S^{inv} be the maps defined in Section 3.6.3. Define $\Phi^*(t)$ as $\Phi^*(t) = \Phi(\pi_S^{inv}(t))$ and let Y_0 be a measure zero set as in Lemma 3.6.6. We claim that $\Phi^*(t)$ satisfies (4.3) for all $t \in [0, \pi(S)] \setminus Y_0$. Fix $t \in [0, \pi(S)] \setminus Y_0$. Write

$$\Phi^*(t+r) = \Phi(\pi_S^{inv}(t+r)) \geq \Phi(\pi_S^{inv}(t) + r).$$

Here, we used that (a) $\pi_S^{inv}(t+r) \geq \pi_S^{inv}(t) + r$ and (b) Φ is a monotone function. By Lemma 3.6.6, $\pi_S^{inv}(t) \in S$, thus

$$\Phi^*(t+r) \geq \Phi(\pi_S^{inv}(t) + r) \geq \Phi(\pi_S^{inv}(t)) + \gamma \int_0^{\pi_S^{inv}(t)} \Phi(x)dx.$$

We now observe that $\Phi^*(t) = \Phi(\pi_S^{inv}(t))$ and

$$\begin{aligned}
\int_0^{\pi_S^{inv}(t)} \Phi(x)dx &\geq \int_0^{\pi_S^{inv}(t)} \Phi(x) \cdot \mathbb{1}(x \in S)dx \\
&= \int_0^{\pi_S^{inv}(t)} \Phi(x)d\pi_S(x) \\
&= \int_0^t \Phi^*(x)dx.
\end{aligned}$$

Here, we used that $d\pi_S(x) = \mathbb{1}(x \in S)dx$ and $\pi_S(\pi_S^{inv}(t)) = t$. Thus, we showed that for all $t \in [0, \pi(S)] \setminus Y_0$, we have

$$\Phi^*(t+r) \geq \Phi^*(t) + \int_0^t \Phi^*(x)dx.$$

We now use Lemma 4.2.1 with function Φ^* and $R' = \mu(S)$. We obtain the following inequality:

$$\Phi^*(\mu(S)) \geq e^{\eta\mu(S)-1},$$

which concludes the proof of Lemma 3.6.6. □

REFERENCES

[1] S. Ahmadi, S. Khuller, and B. Saha. Min-max correlation clustering via multicut. In *Proceedings of the Conference on Integer Programming and Combinatorial Optimization*, pages 13–26, 2019.

[2] K. Ahn, G. Cormode, S. Guha, A. McGregor, and A. Wirth. Correlation clustering in data streams. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2237–2246, 2015.

[3] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

[4] N. Ailon, N. Avigdor-Elgrabli, E. Liberty, and A. van Zuylen. Improved approximation algorithms for bipartite correlation clustering. *SIAM Journal on Computing*, 41(5): 1110–1121, 2012.

[5] N. Ailon, Y. Chen, and H. Xu. Breaking the small cluster barrier of graph clustering. In *International Conference on Machine Learning*, pages 995–1003, 2013.

[6] N. Amit. The bicluster graph editing problem. *Tel Aviv University*, 2004.

[7] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine learning*, 56(1-3): 89–113, 2004.

[8] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 184–193. IEEE, 1996.

[9] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.

[10] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference*, pages 595–601, 2004.

[11] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711–726, 2004.

[12] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the International Conference on World Wide Web*, pages 587–596, 2011.

[13] P. Boldi, A. Marino, M. Santini, and S. Vigna. BUbiNG: Massive crawling for the masses. In *Proceedings of the Companion Publication of the International Conference on World Wide Web*, pages 227–228, 2014.

[14] F. Bonchi, D. García-Soriano, and E. Liberty. Correlation clustering: from theory to practice. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1972, 2014.

[15] G. Călinescu, H. Karloff, and Y. Rabani. An improved approximation algorithm for multiway cut. *Journal of Computer and System Sciences*, 60(3):564–574, 2000.

[16] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *IEEE Symposium on Foundations of Computer Science*. Citeseer, 2003.

[17] M. Charikar, N. Gupta, and R. Schwartz. Local guarantees in graph cuts and clustering. In *Proceedings of the Conference on Integer Programming and Combinatorial Optimization*, pages 136–147, 2017.

[18] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *Computational Complexity*, 15(2): 94–114, 2006. doi: 10.1007/s00037-006-0210-9. URL `https://doi.org/10.1007/s00037-006-0210-9`.

[19] S. Chawla, K. Makarychev, T. Schramm, and G. Yaroslavtsev. Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *Proceedings of the Symposium on Theory of Computing*, pages 219–228, 2015.

[20] F. Chierichetti, N. Dalvi, and R. Kumar. Correlation clustering in mapreduce. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 641–650, 2014.

[21] W. Cohen and J. Richman. Learning to match and cluster entity names. In *Proceedings of the ACM SIGIR-2001 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2001.

[22] W. W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 475–480, 2002.

[23] V. Cohen-Addad, S. Lattanzi, S. Mitrovic, A. Norouzi-Fard, N. Parotsidis, and J. Tarnawski. Correlation clustering in constant many parallel rounds. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2069–2078, 2021.

[24] V. Cohen-Addad, C. Fan, S. Lattanzi, S. Mitrovic, A. Norouzi-Fard, N. Parotsidis, and J. Tarnawski. Near-optimal correlation clustering with privacy. *CoRR*, abs/2203.01440, 2022. doi: 10.48550/arXiv.2203.01440. URL `https://doi.org/10.48550/arXiv.2203.01440`.

[25] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006.

[26] M. Elsner and W. Schudy. Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 19–27. Association for Computational Linguistics, 2009.

[27] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497, 2004.

[28] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 418–426, 2003.

[29] N. Garg, V. V. Vazirani, and M. Yannakakis. Approximate max-flow min-(multi) cut theorems and their applications. *SIAM Journal on Computing*, 25(2):235–251, 1996.

[30] I. Giotis and V. Guruswami. Correlation clustering with a fixed number of clusters. *CoRR*, abs/cs/0504023, 2005. URL `http://arxiv.org/abs/cs/0504023`.

[31] J. Jafarov, S. Kalhan, K. Makarychev, and Y. Makarychev. Correlation clustering with asymmetric classification errors. In *Proceedings of the International Conference on Machine Learning*, pages 4641–4650, 2020.

[32] J. Jafarov, S. Kalhan, K. Makarychev, and Y. Makarychev. Local correlation clustering with asymmetric classification errors. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4677–4686. PMLR, 18–24 Jul 2021.

[33] S. Kalhan, K. Makarychev, and T. Zhou. Correlation clustering with local objectives. In *Advances in Neural Information Processing System*, pages 9341–9350, 2019.

[34] S. Lattanzi, B. Moseley, S. Vassilvitskii, Y. Wang, and R. Zhou. Robust online correlation clustering. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[35] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan. Correlation clustering with noisy partial information. In *Conference on Learning Theory*, pages 1321–1342, 2015.

[36] A. W. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families I: Bipartite Ramanujan graphs of all degrees. In *Proceedings of the Symposium on Foundations of Computer Science*, pages 529–537, 2013.

[37] C. Mathieu and W. Schudy. Correlation clustering with noisy input. In *Proceedings of the Symposium on Discrete Algorithms*, pages 712–728, 2010.

[38] C. Mathieu, O. Sankur, and W. Schudy. Online Correlation Clustering. In J.-Y. Marion and T. Schwentick, editors, *27th International Symposium on Theoretical Aspects of Computer Science*, volume 5 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 573–584, Dagstuhl, Germany, 2010. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-16-3. doi: 10.4230/LIPIcs.STACS.2010.2486. URL `http://drops.dagstuhl.de/opus/volltexte/2010/2486`.

[39] X. Pan, D. Papailiopoulos, S. Oymak, B. Recht, K. Ramchandran, and M. I. Jordan. Parallel correlation clustering on big graphs. In *Advances in Neural Information Processing Systems*, pages 82–90, 2015.

[40] G. J. Puleo and O. Milenkovic. Correlation clustering and biclustering with locally bounded errors. *IEEE Transactions on Information Theory*, 64(6):4105–4119, 2018.

[41] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proceedings of the Conference on Computer and Communications Security*, pages 342–351, 2007.

[42] C. Swamy. Correlation clustering: Maximizing agreements via semidefinite programming. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, page 526–527, 2004.

[43] J. Tan. A note on the inapproximability of correlation clustering. *ArXiv*, abs/0704.2092, 2008.

[44] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*, pages 100–111, 2016.

[45] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017.

[46] J. Van Gael and X. Zhu. Correlation clustering for crosslingual link detection. In *IJCAI*, pages 1744–1749, 2007.

[47] A. Wirth. Correlation clustering. In *Encyclopedia of Machine Learning*, pages 227–231. Springer, 2010.