Integer Programming Models and Polyhedral Study for the Geodesic Classification Problem on Graphs

Paulo H. M. Araújo Manoel Campêlo Ricardo C. Corrêa Martine Labbé

July 19, 2022

Abstract

We study a discrete version of the classical classification problem in the Euclidean space, to be called geodesic classification problem. It is defined on a graph, where some vertices are initially assigned a class and the remaining ones must be classified. This vertex partition into classes is grounded on the concept of geodesic convexity on graphs, as a replacement for the Euclidean convexity in the multidimensional space. We propose two new integer programming models along with branch-and-cut algorithms to solve them. We also carry out a polyhedral study of the associated polyhedra, which includes families of facet-defining inequalities and separation algorithms. Finally, we run computational experiments to evaluate the computational efficiency and the classification accuracy of the proposed approaches by comparing them with classic solution methods for the Euclidean convexity classification problem.

Keywords: classification. geodesic convexity. polyhedral combinatorics.

1 Introduction

Supervised learning stands for an automatic prediction tool widely used in many situations in nowadays information society. In general terms, it denotes a collection of methods that act on partial information in order to infer the structure of the entire universe with respect to a specific target property. Most frequently, the partial information provided consists of a set of samples whose target property assignments are known, as well as some relationship between these samples. In this context, the automatic prediction is performed through the following two-phase procedure: in the initial phase, or *training phase*, a given sample set is analyzed. Each sample consists of an array of encoded attributes that characterize an object of a certain type together with a label that associates a class to the corresponding object. Most commonly, the target property is the partition of the universe of possible objects in two classes. A tacit assumption made at this phase is that there is an underlying pattern associated with the samples of each class that sets them apart from the samples of the other classes. Thus, the purpose of the training phase is to determine a mapping from all possible objects of the considered type into the set of possible classes as an extension of an underlying pattern of the samples. Then, in the second phase, the mapping determined in the training phase is used to respond to queries about the class of objects that do not belong to the sample set.

An optimization problem is usually associated with the training phase. Referred to as *classification problem*, it consists in grouping similar samples to get clusters as internally homogeneous as possible. A wide range of solution methods is available, each depending on the coding of the samples and the criterion adopted to express homogeneity. A prevalent approach, which we call *Euclidean classification*, is to encode the samples as vectors of numerical features in a multidimensional Euclidean space and to assume that the class patterns can be appropriately characterized by convex sets. More precisely, consider that the samples are colored points in $\mathbb{R}^d$, for some $d \geq 1$, the color representing the class of the sample. The goal is to assign a class (color) to every point in $\mathbb{R}^d$ based on the classification of the samples so that the convex hulls of the colors do not intersect (possibly disregarding some samples as discussed below). In this vein, continuous optimization methods, including linear and quadratic programming, have been developed in the last 40 years. See e.g. [12, 21]. More recently, integer linear programming tools started to be used in conjunction with continuous methods, as we can see in [5, 24].

In [1], a new variant of the classification problem was defined. For this purpose, a correspondence between the convexity concepts in discrete and continuous mathematics can be established if we consider the vertex set of a connected graph and the distance between vertices as metric space. Thus, the *geodesic classification* problem is stated in terms of notions of convexity in graphs and assumes the following hypotheses:

1. The universe of objects is a discrete set in which each object is not characterized by its features (which are not necessarily numerical) but, instead, by its similarities with other objects. The configuration of the objects is thus represented by a similarity graph $G = (V, E)$, connected, where V is the set of all objects, and E gives the pairs of similar objects. Usually, the definition of E depends on the features of the objects. The vertices associated with the sample set constitute a proper subset of V.

2. There exists an underlying classes pattern that can be expressed, or at least approximated, by the notion of *geodesic convexity* in graphs [22]. Such a convexity is defined with respect to the shortest paths in G (analogously to the definition of Euclidean convexity with respect to the Euclidean distances between points in $\mathbb{R}^n$). In addition, it is assumed that the shape of the classes among the samples spans the classes patterns over the graph according to geodesic convexity.

3. The sample set may contain an arbitrary number of misclassified objects, called *outliers*, which result from possible sampling errors or due to inherent characteristics of the phenomenon being modeled. From the mathematical point of view, an outlier is a classified object that leads the underlying pattern of the samples in its class to deviate from the convexity definition. The possible occurrence of outliers poses an additional challenge to any method used to solve the classification problem since they have to be detected and disregarded so that an accurate solution may be found.

The goal is to split the vertex set into classes, based on the classification of the samples and the structure of the similarity graph, in such a way that an error measure in a metric space is minimized [2]. In this paper, we consider the existence of only two classes in the graph and the

number of disregarded outliers as the error measure. The classification of the vertices follows a specific notion of *linear separability* with respect to geodesic convexity.

From the practical point of view, this problem allows encoding object similarities through some reflexive binary relation. This fact benefits many practical applications in big data, specially in two situations that can arise even when objects are modeled as points in an Euclidean space. The first situation occurs when similarities are expressed in terms of symmetric and non-transitive binary relations. Such a relation define an unweighted similarity graph G. A standard example is to consider as similar any two points that are close to each other in an Euclidean space. The second situation, which arises very often when handling multiple models of text corpora, are constituted by symmetric and transitive relations, thus leading G to be a complete graph (recall that G is assumed to be connected). The particularity of this case is that G is edge-weighted, the weight of an edge standing for a degree of similarity between objects. Cosine similarity in text analysis is an example (see [13] for a general tool based on two distinct topic modeling methods). The theoretical results discussed in this paper assume the first type of similarity relation, but they can be extended directly to the second type if the edge weights are considered in the definition of path length and, consequently, of geodesic convexity. Due to its characteristics, applications of the geodesic classification problem are easily found in the fields of data mining and classical statistics. Text and sentiment analyses, community detection in complex networks (such as social networks and networks of citations of scientific articles), historic files similarity prediction, content recommendation in video streaming services, and spam filtering for e-mails constitute examples thereof [15].

Geodesic and Euclidean classification are distinct problems in the following sense. Consider a set V of points in a multidimensional Euclidean space, and the corresponding similarity graph $G = (V, E)$ such that E connects points whose Euclidean distance is smaller than a given threshold. Assuming, as mentioned above, that the samples form a proper subset of V, the Euclidean classification problem consists in partitioning the Euclidean space into two convex subspaces based on the classification of the samples. Although based on the same samples, the geodesic classification problem aims to split set V only. It is shown in details in Section 2 that there are feasible solutions in the geodesic problem where the classes assigned to points in V that are not considered as outliers make the Euclidean problem infeasible. The converse is also true. Put differently, the possible patterns considered in each problem are distinct. The main reason is that the universe in the former case is composed of selected points of the Euclidean space only. Actually, a solution for the geodesic problem is neither a covering nor a partitioning of G in convex sets in the sense studied in [3] and [8]. Instead, this problem can be seen as the combination of a graph convexity problem and the well-known set covering problem [16], as shown by the mathematical model proposed in Section 3.

We have introduced the 2-class geodesic classification problem (2-GC) in the conference paper [1], where preliminary results were presented, including a integer formulation. In this work, we present two new integer formulations for 2-GC along with a branch-and-cut algorithm for each one. Besides, we run several computational experiments with random and realistic instances to evaluate the geodesic convexity approach. The first formulation has a linear number of variables but an exponential number of constraints, whereas the second one is an extended formulation

with more variables but a polynomial number of constraints. An interesting feature of the first model is that it expresses the 2-GC problem as a set covering problem. Thus, we can take advantage of well-known results from the literature.

We also study the polytopes associated with each formulation. We show that one of them is an orthogonal projection of the other. Most of the derived facet-defining inequalities can be seen as counterparts of those presented by [11] for a polyhedron that models the Euclidean version of the problem. On the other hand, some of them originates from specific properties of the geodesic case. Despite the fact that the geodesic and Euclidean classification problems are distinct, the study of the combinatorial structure of the former may be useful to design solution methods for the latter. Actually, the difference between the problems also justifies and motivates the investigation of 2-GC.

This paper is organized as follows. The formal definition of the geodesic classification problem is introduced in Section 2, and the integer linear formulations with the associated polyhedra are studied in Sections 3 and 4. In Section 5, we present the branch-and-cut algorithms and evaluate the performance of each approach through computational experiments. We use randomly generated instances as well as instances derived from real applications. Some of them are obtained from instances for the Euclidean case, so that we could compare the accuracy of our approaches with well-known methods from the literature. Finally, we present concluding remarks and directions for future works in Section 6. Basic concepts, notation, and results of graph theory, linear algebra, the set covering problem, polyhedra, linear programming, duality, *branch-and-bound* and *branch-and-cut* can be found in [20].

2 Geodesic Classification Problem

In this section, we formalize the geodesic classification problem as introduced in [1].

2.1 Linear Separability

We follow the basic concepts and terminology in graph theory adopted in [7]. In special, a *path (between two vertices u and v)* in a graph $G = (V, E)$ is $\langle u \rangle$, if $u = v$, or a sequence of distinct vertices $\langle u = v_1, v_2, \ldots, v_\ell = v \rangle$, $\ell \geq 1$, such that $\{v_i, v_{i+1}\} \in E$ for $i \in \{1, \ldots, \ell - 1\}$. The *length* of the path is the number of its vertices minus 1, that is, $\ell - 1$. When $\ell - 1$ is the minimum length among all paths between u and v, then $\delta(u, v) = \ell - 1$ and $\langle u = v_1, v_2, \ldots, v_\ell = v \rangle$ defines a *shortest path* or *geodesic* between u and v. The operator D applied to $S \subseteq V$ gives the set $D[S]$ of all vertices lying on any geodesic between pairs of vertices in S, *i.e.* $D[S] = \{w \mid$ there exist $u, v \in S$ and a geodesic between u and v in G containing $w\}$. If $S = \{u, v\}$, then we write $D[u, v] = D[\{u, v\}]$ and $D(u, v) = D[u, v] \setminus \{u, v\}$. Notice that $D[S] = \bigcup_{u,v \in S} D[u, v]$. If $D[S] = S$, then S is a *convex set*. The *convex hull* of S, denoted by $H[S]$ (and by $H[u, v]$ if $S = \{u, v\}$), is the minimal convex set containing S. This minimum set is unique. Observe that $H[S] = D^k[S]$ for some $k \geq 1$, which may lead to $H[u, v] \neq D[u, v]$ in some cases. In other terms, $H[S]$ can be obtained by the iterative application of D. The subset S is called a *basis* of $H[S]$, which *spans* on G through D to express $H[S]$. For more details about convexity on graphs, we refer to [3].

For the purpose of formally defining the geodesic classification problem, let $G = (V = V_B \cup V_R \cup V_N, E)$ be a similarity graph where $V_B, V_R \subseteq V$ are disjoint subsets of, respectively, *blue* and *red* classes. These are initially classified vertices and form the set $V_{BR} = V_B \cup V_R$, $|V_{BR}| = n$, of samples. The remaining vertices in V are the unclassified (*neutral*) vertices and define the set V_N. Whenever we refer to a non-specific color $K \in \{B, K\}$, let $\bar{K}$ denote the opposite color. The first step to state the problem is the notion of *linear separability*, as follows.

Definition 1 (Linear separability [1])**.** *A pair* $(A_B \subseteq V_B, A_R \subseteq V_R)$ *is* linearly separable (in G with respect to V_B, V_R, V_N) *if*

(C1) $H[A_B] \cap A_R = \emptyset$*,*

(C2) $H[A_R] \cap A_B = \emptyset$*, and*

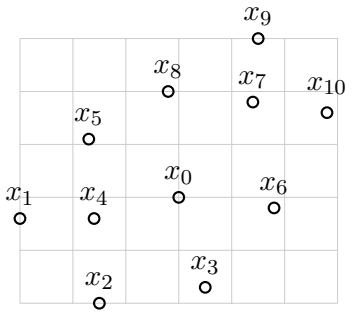
(C3) $H[A_B] \cap H[A_R] \cap V_N = \emptyset$

hold, and is linearly inseparable *otherwise.*

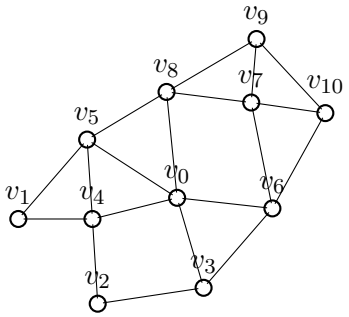
We observe that this definition is not equivalent to the one usually adopted in Euclidean classification. For the sake of comparison, let us consider the situation depicted in Figure 1 which shows a set of points $X = \{x_0, \ldots, x_{10}\}$ in the plane. The usual scenario in the case consists in sampling finite sets X_R and X_B of red and blue points, respectively, in order to infer the class of all points in $\mathbb{R}^2$. A more appropriate scenario to establish a parallel with the geodesic version occurs when only the class of a finite subset X_N of $\mathbb{R}^2 \setminus (X_R \cup X_B)$ is required. Usually, the set X_N is large and not known in advance. In the example shown in Figure 1a, let us consider $X_R = \{x_2, x_3, x_4\}$, $X_B = \{x_6, x_7, x_8\}$, and $X_N = \{x_0, x_1, x_5, x_9, x_{10}\}$. The usual approach is to associate the notion of linear separability in the Euclidean case with a hyperplane that separates X_R from X_B, which corresponds to have $\text{conv}(X_B) \cap \text{conv}(X_R) = \emptyset$, where conv denotes the convex hull in a Euclidean space [11]. Clearly, X_R and X_B are linear separable in this example.

Figure 1 – Points in a bidimensional Euclidean space and a similarity graph.

(a) Points in $\mathbb{R}^2$.



(b) Points as close as x_6 and x_7, or closer, are similar.

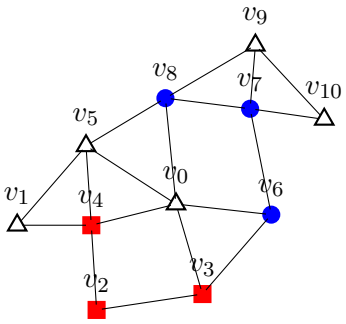


In Figure 1b, the similarity graph is obtained using Euclidean distances between the points shown in Figure 1a. The pair (V_B, V_R) shown in Figure 2a is the counterpart of (X_B, X_R). Although (X_B, X_R) is linearly separable in Euclidean space, (V_B, V_R) is linearly inseparable in the similarity graph. Indeed, v_0 is simultaneously in a geodesic between two red vertices (v_3, v_4) and two blue vertices (v_6, v_8). Thus, $v_0 \in H[V_B] \cap H[V_R] \cap V_N$ so that Condition (C3) is violated.

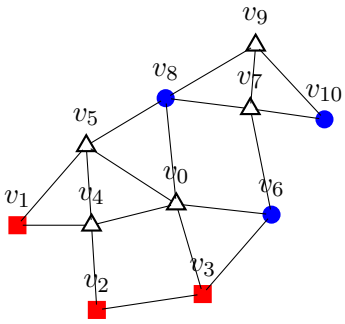
However, $(V_B \setminus \{v_i\}, V_R)$, for any $i \in \{6, 8\}$, as well as $(V_B, V_R \setminus \{v_i\})$, for any $i \in \{3, 4\}$, are linearly separable. Similarly, (V_B, V_R) of Figure 2b is also linearly inseparable.

Figure 2 – Similarity graph of Figure 1b with two linearly inseparable pairs (V_B, V_R) such that $(V_B \setminus \{v_6\}, V_R)$ is linearly separable. In both cases, filled circles are vertices of V_B, filled squares represent V_R, and unfilled triangles represent V_N.

(a) The red basis $A_R = \{v_2, v_3, v_4\}$ spans $H[A_R] = A_R \cup \{v_0\}$ and the blue basis $A_B = \{v_7, v_8\}$ spans $H[A_B] = A_B$.



(b) The red basis $A_R = \{v_2, v_3, v_4\}$ spans $H[A_R] = A_R \cup \{v_0, v_4, v_5\}$ and the blue basis $A_B = \{v_8, v_{10}\}$ spans $H[A_B] = A_B \cup \{v_7, v_9\}$.



2.2 Problem Statement

In terms of the linear separability defined above, a pair $(A_B \subseteq V_B, A_R \subseteq V_R)$ being linearly separable means taking $V_B \setminus A_B$ and $V_R \setminus A_R$ as outliers. In other words, an *outlier* is a vertex $v \in V_B \cup V_R$ whose color is disregarded when calculating the convex hull of the red and blue bases. However, it is worth remarking that it is not removed from the similarity graph and is still considered for determining the shortest paths. With this in mind, we can define the geodesic classification problem associated with G, V_B, V_R, and V_N as the determination of a linearly separable pair that optimizes some linear function. We postpone the discussion on the objective function until Section 5.

Problem (2-Class Geodesic Classification (2-GC))**.** *Given a connected graph* $G = (V = V_B \cup V_R \cup V_N, E)$*, find linearly separable subsets* $(A_B \subseteq V_B,\ A_R \subseteq V_R)$ *that optimizes some linear function on* G*,* A_B*, and* A_R*.*

Any feasible solution (A_B, A_R) of the 2-GC problem defines a mapping from $H[A_B] \cup H[A_R]$ onto $\{blue, red\}$, which classifies all neutral vertices in $H[A_B]$ as blue and those in $H[A_R]$ as red. Note that, as illustrated in Figure 2a, $V_N \setminus (H[A_B] \cup H[A_R])$ can be nonempty. These vertices can be independently assigned to either of the classes without violating the linear separability or increasing the the number of outliers. Not suprisingly, a similar phenomenon occurs when we deal with the concept of linear separation in the Euclidean classification. Since there are many separating hyperplanes, there exist some regions of the space that can be assigned distinct classes depending on the feasible solution considered. In particular, this is the case of points x_0 and x_5 in Figure 1a. Solution methods seek to select the hyperplane that best segregates the two classes according to some additional criteria (for example, by maximizing the distances between the closest data points of either class and the hyperplane).

Two final remarks are worthwhile in connection with the geodesic classification. First, we observe that the 2-GC problem always has a solution since we could consider all initially classified vertices of a class as outliers. Second, due to the outliers and the vertices in $V_N \setminus (H[A_B] \cup H[A_R])$, each class does not necessarily define a convex set, and $(H[A_B], H[A_R])$ is neither a covering nor a packing of the vertices of G.

3 A Set Covering Formulation for the 2-GC Problem

3.1 Integer Formulation

In this section, we formulate the 2-GC problem as a set covering problem of the form

$$\min \left\{ \mathbf{1}^\top \boldsymbol{y} \mid A\boldsymbol{y} \geq \mathbf{1}, \boldsymbol{y} \in \mathbb{B}^n \right\}, \tag{1}$$

where A is a 0-1 matrix and $\mathbf{1}$ is the vector of ones. For a general binary matrix A, (1) is NP-hard [16] and results about valid inequalities and facet-defining properties can be found in [23]. In the particular case of the 2-GC problem, we use a binary variable y_i, for each vertex $i \in V_{BR}$, such that $y_i = 1$ if i is an outlier, and $y_i = 0$ otherwise. Then, using $K(i) \in \{B, R\}$ and $\bar{K}(i) \in \{B, R\} \setminus \{K(i)\}$ to respectively denote the class and the opposite class of vertex $i \in V_{BR}$, matrix A corresponds to the following constraints:

$$\sum_{j \in S \cup \{i\}} y_j \geq 1, \qquad i \in V_{BR}, S \subseteq V_{\bar{K}(i)} : i \in H[S], \tag{2}$$

$$\sum_{j \in S \cup T} y_j \geq 1, \qquad S \subseteq V_B, T \subseteq V_R : H[S] \cap H[T] \cap V_N \neq \emptyset. \tag{3}$$

This formulation will be called ILP1.

Proposition 2. *Inequalities* (2) *and* (3) *define the feasible solutions of the 2-GC problem.*

Proof. Let $\boldsymbol{y} \in \mathbb{B}^n$ satisfies (2) and (3). Define $A_K = \{i \in V_K : y_i = 0\}$, for $K \in \{B, R\}$. We want to show that (A_B, A_R) satisfies (C1)-(C3). Suppose that Condition (C1) is violated and let $i \in A_R \cap H[A_B]$. By definition, $y_j = 0$ for all $j \in A_B \cup \{i\}$, which violates (2) for i and $S = A_B$: a contradiction. A similar contradiction is obtained by assuming that Condition (C2) does not hold. Besides, supposing Condition (C3) violated contradicts (3) for $S = A_B$ and $T = A_R$. Therefore, (C1)-(C3) are satisfied.

Conversely, let $A_B \subseteq V_B$ and $A_R \subseteq V_R$ satisfying (C1)-(C3). Define $\boldsymbol{y} \in \mathbb{B}^n$ such that $y_i = 0$ if $i \in A_B \cup A_R$ and $y_i = 1$ otherwise. We first consider constraints (2). Let $i \in V_R$ and $S \subseteq V_B$ such that $i \in H[S]$. If $S \subseteq A_B$, Condition (C1) implies $i \in V_R \setminus A_R$. Otherwise, there exists $j \in S$ such that $j \in V_B \setminus A_B$. In both cases, (2) for i and S is satisfied. Using Condition (C2), we get a similar result for $i \in V_B$ and $S \subseteq V_R$. Finally, let $S \subseteq V_B$ and $T \subseteq V_R$ such that $H[S] \cap H[T] \cap V_N \neq \emptyset$ defining a constraint (3). It follows from Condition (C3) that $S \setminus A_B \neq \emptyset$ or $T \setminus A_R \neq \emptyset$. Therefore, there exists $j \in S \cup T$ that implies (3) for S and T. □

Observe that $(S, \{i\})$ (or $(\{i\}, S)$) is linearly inseparable in (2) whereas (S, T) is linearly

inseparable in (3). So, inequalities (2)-(3) are special cases of the valid inequalities

$$\sum_{j \in S \cup T} y_j \geq 1, \quad (S,T) \text{ linearly inseparable.} \tag{4}$$

A linearly inseparable pair (S,T) is said to be *minimal* if $(S \setminus \{u\}, T \setminus \{u\})$, for every $u \in S \cup T$, is linearly separable. Since $S \cap T = \emptyset$, note that either $S \setminus \{u\} = S$ or $T \setminus \{u\} = T$. An inequality in (4) is clearly redundant when (S,T) is not minimal, and this particularly applies to constraints (2)-(3). Actually, in Subsection 3.4 we derive necessary and sufficient facetness conditions for these constraints. By now, we characterize minimal linearly inseparable pairs.

Proposition 3. *A linearly inseparable pair (S,T) is minimal if, and only if, one of the following conditions holds: (i) $S = \{i\}$ and $i \notin H[T \setminus \{u\}]$, for all $u \in T$; or (ii) $T = \{i\}$ and $i \notin H[S \setminus \{u\}]$, for all $u \in S$; or (iii) $H[S] \cap T = H[T] \cap S = \emptyset$, $H[S \setminus \{u\}] \cap H[T \setminus \{u\}] \cap V_N = \emptyset$, for all $u \in S \cup T$.*

Proof. First, assume that (i) holds. Note that $H[\emptyset] = \emptyset$ and $H[i] = \{i\}$. Then, $(S \setminus \{u\}, T \setminus \{u\})$, for every $u \in \{i\} \cup T$, trivially satisfies conditions (C1)-(C3). This means that $(S \setminus \{u\}, T \setminus \{u\})$ is linearly separable and so (S,T) is minimal. A similar result is attained under condition (ii). Now, assume condition (iii) and let $u \in S \cup T$. Then, $H[S \setminus \{u\}] \cap (T \setminus \{u\}) \subseteq H[S] \cap T = \emptyset$, $H[T \setminus \{u\}] \cap (S \setminus \{u\}) \subseteq H[T] \cap S = \emptyset$, and $H[S \setminus \{u\}] \cap H[T \setminus \{u\}] \cap V_N = \emptyset$. Therefore, $(S \setminus \{u\}, T \setminus \{u\})$ is linearly separable and so (S,T) is minimal.

Conversely, assume that (S,T) is minimal. Then, for every $u \in S \cup T$, $H[S \setminus \{u\}] \cap (T \setminus \{u\}) = \emptyset$, $H[T \setminus \{u\}] \cap (S \setminus \{u\}) = \emptyset$, and $H[S \setminus \{u\}] \cap H[T \setminus \{u\}] \cap V_N = \emptyset$. Suppose that (iii) does not hold, which implies $H[S] \cap T \neq \emptyset$ or $H[T] \cap S \neq \emptyset$. It remains to show that (i) or (ii) are satisfied. First, assume $H[S] \cap T \neq \emptyset$. Then, $|S| \geq 2$ and there is $i \in T$ such that $i \in H[S]$. It follows that $(S, \{i\})$ is linearly inseparable. Since (S,T) is minimal, it must be $T = \{i\}$ and $i \notin H[S \setminus \{u\}]$ for every $u \in S$. Therefore, (ii) holds. Similarly, if $H[T] \cap S \neq \emptyset$, we can conclude that condition (i) holds. □

3.2 Polyhedra

We turn our attention in this section to basic properties of the polyhedron associated with the integer formulation, defined as $P_1 = \text{conv}\{\boldsymbol{y} \in \mathbb{B}^n \mid (2)\text{–}(3)\}$, where $n = |V_{BR}|$. The first result stems from the fact that each constraint of the integer formulation contains at least two nonnull coefficients [4]. For convenience, we present an alternative proof introducing some elements and giving better intuition on the facetness conditions to be discussed later. Besides using $\mathbf{0}$ and $\mathbf{1}$ to denote the null vector and the vector of ones, respectively, we adopt the notation $\boldsymbol{e}^i$ to represent the i-th unit vector and $\bar{\boldsymbol{e}}^i = \mathbf{1} - \boldsymbol{e}^i$.

Proposition 4. *P_1 is full-dimensional.*

Proof. Consider the $n+1$ affinely independent points $\mathbf{1}$ and $\bar{\boldsymbol{e}}^i$, for all $i \in V_{BR}$. To show that they are in P_1, first note that $\mathbf{1}$ satisfies all inequalities (2) and (3). Let $i \in V_{BR}$ and $S \subseteq V_{\bar{K}(i)}$ such that $i \in H[S]$. By definition, S is nonempty and $i \notin S$. Then, there is at least one vertex $j \in S$ such that $j \neq i$, and so $\bar{e}^i_j = 1$, which implies that $\bar{\boldsymbol{e}}^i$ satisfies (2) for i and S. This

inequality is also satisfied by $\bar{\boldsymbol{e}}^{i'}$, $i' \neq i$, since $\bar{e}_i^{i'} = 1$. Since every inequality in (3) involve two disjoint and non-empty subsets S and T, there is at least one vertex $j \in S \cup T$ such that $j \neq i$, and so $\bar{e}_j^i = 1$, which yields that $\bar{\boldsymbol{e}}^i$ satisfies all inequalities (3). □

The solutions used in the proof of Proposition 4 allow us to show that the bounding inequalities induce facets of P_1.

Proposition 5. *For every $i \in V_{BR}$, $y_i \geq 0$ and $y_i \leq 1$ are facet-defining for P_1.*

Proof. Let $i \in V_{BR}$. The face defined by $y_i \leq 1$ contains the affinely independent points $\mathbf{1}$ and $\bar{\boldsymbol{e}}^j$, for all $j \in V_{BR} \setminus \{i\}$. The face defined by $y_i \geq 0$ contains the affinely independent points $\bar{\boldsymbol{e}}^i$ and $\bar{\boldsymbol{e}}^i - \boldsymbol{e}^j$, for all $j \in V_{BR} \setminus \{i\}$. □

Again, the following result is proved in [4] and we give an alternative proof for convenience.

Proposition 6. *If a facet-defining inequality $\pi^\top \boldsymbol{y} \geq \pi_0$ of P_1 is different from $y_i \leq 1$ and $y_i \geq 0$, for all $i \in V_{BR}$, then $\pi \geq \mathbf{0}$ and $\pi_0 > 0$.*

Proof. Let $i \in V_{BR}$. If $\pi^\top \boldsymbol{y} \geq \pi_0$ is different from $y_i \leq 1$, then the facet $F := \{\boldsymbol{y} \in P_1 \mid \pi^\top \boldsymbol{y} = \pi_0\}$ contains a point $\bar{\boldsymbol{y}}$ with $\bar{y}_i = 0$. Since the point $\boldsymbol{y'} = \bar{\boldsymbol{y}} + \boldsymbol{e}^i$ is also in P_1, we get $\pi^\top \boldsymbol{e}^i = \pi^\top(\boldsymbol{y'} - \bar{\boldsymbol{y}}) \geq \pi_0 - \pi_0 = 0$, which leads to $\pi_i \geq 0$. Considering that $\pi \neq 0$, assume that $\pi_i > 0$. Hence, F contains a point $\hat{\boldsymbol{y}}$ with $\hat{y}_i = 1$ when the inequality is different from $y_i \geq 0$. Then, $\pi_0 = \pi^\top \hat{\boldsymbol{y}} \geq \pi_i > 0$ in this case. □

3.3 $\mathcal{N}$-Set Inequalities

Let $(S \subseteq V_B, T \subseteq V_R)$ be linearly inseparable. Analogously to the definition in [11] for the Euclidean case, let an *$\mathcal{N}$-set for (S,T)* be a minimal set $N \subseteq S \cup T$ such that $(S \backslash N, T \backslash N)$ is linearly separable. We define $\mathcal{N}(S,T) = \{N \subseteq S \cup T \mid N \text{ is an } \mathcal{N}\text{-set for } (S,T)\}$, and for each $i \in S \cup T$,

$$\nu_i = \min\{|N| \mid N \in \mathcal{N}(S,T), i \in N\}.$$

We assume that $\nu_i = \infty$ if $\{N \in \mathcal{N}(S,T) \mid i \in N\} = \emptyset$. Observe that $\nu_i \leq |N|$, for all $N \in \mathcal{N}(S,T)$ and $i \in N$.

Proposition 7. *Let $(S \subseteq V_B, T \subseteq V_R)$ be linearly inseparable. The* $\mathcal{N}$-set inequality

$$\sum_{i \in S \cup T} \frac{y_i}{\nu_i} \geq 1 \tag{5}$$

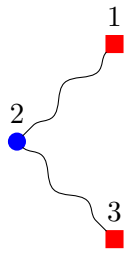
is valid for P_1.

Proof. Let $\bar{\boldsymbol{y}}$ be a feasible solution and define the nonempty set $N' = \{i \in S \cup T \mid \bar{y}_i = 1\}$. Then, there exists $N \subseteq N'$ such that $N \in \mathcal{N}(S,T)$. This leads to $\sum_{i \in S \cup T} \frac{\bar{y}_i}{\nu_i} = \sum_{i \in N'} \frac{1}{\nu_i} \geq \sum_{i \in N} \frac{1}{\nu_i} \geq \sum_{i \in N} \frac{1}{|N|} = 1$, where the equality and first inequality follow from the definition of N' and $N \subseteq N'$, respectively, whereas the last inequality is due to $N \in \mathcal{N}(S,T)$, which implies $\nu_i \leq |N|$, for all $i \in N$. □

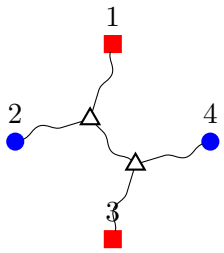
Some special cases of (5) are illustrated in Figure 3. We assume hereafter, without loss of generality, that ν_i is finite for all $i \in S \cup T$. Otherwise, $(S \setminus I, T \setminus I)$, for $I = \{i \in S \cup T \mid \nu_i = \infty\}$, is also linearly inseparable and defines the same inequality. If $(S \setminus I, T \setminus I)$ were linearly separable, there would be an $\mathcal{N}$-set $N \subseteq I \subseteq (S \cup T)$ and we would have ν_i finite for all $i \in N \subseteq I$.

Figure 3 – Examples of $\mathcal{N}$-set inequalities. Facet defining structures according to Theorem 9. S and T given by blue circles and red squares, respectively. Snaked segments represent shortest paths between their endpoints.

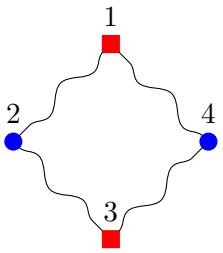
(a) $\sum_{i=1}^{3} y_i \geq 1$. Safe graph is complete. Instance of (2) and Corollary 11.



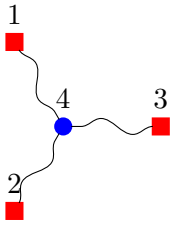
(b) $\sum_{i=1}^{4} y_i \geq 1$. Safe graph is complete. Instance of (3) and and Corollary 11.



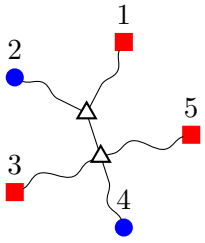
(c) $\sum_{i=1}^{4} y_i \geq 2$. Safe graph is complete. See Proposition 12.



(d) $y_1 + y_2 + y_3 + 2y_4 \geq 2$. Safe graph is complete. Instance of Proposition 13.



(e) $y_1 + 2y_2 + y_3 + 2y_4 + y_5 \geq 2$. Instance of Proposition 14



For $N \subseteq S \cup T$, let $\boldsymbol{y}^N$ be defined as $y_i^N = 0$ if $i \in (S \cup T) \setminus N$, and $y_i^N = 1$ otherwise. An $\mathcal{N}$-set N for (S, T) is *perfect* if $\nu_i = |N|$ for all $i \in N$. Define $\mathcal{N}^*(S, T) = \{N \mid N \text{ is a perfect } \mathcal{N}\text{-set for } (S, T)\}$.

Proposition 8. *If $N \subseteq S \cup T$ is a perfect $\mathcal{N}$-set, then $\boldsymbol{y}^N$ is a point in P_1 satisfying (5) at equality.*

Proof. If N is an $\mathcal{N}$-set, then $(S \setminus N, T \setminus N)$ is linearly separable and so $\boldsymbol{y}^N \in P_1$. In addition, if $\nu_i = |N|$ for all $i \in N$, then $\sum_{i \in S \cup T} \frac{y_i^N}{\nu_i} = \sum_{i \in N} \frac{1}{|N|} = 1$. □

The following notion was introduced in [9, 11]. For each $k \in \mathbb{N}$, the *safe graph* $G_{S,T}^k = (V_{S,T}^k, E_{S,T}^k)$ is defined by $V_{S,T}^k = \{i \in S \cup T \mid \nu_i = k\}$ and $E_{S,T}^k = \{ij \mid \exists N_i, N_j \in \mathcal{N}^*(S, T), i \in N_i, j \in N_j, N_i \ominus N_j = \{i, j\}\}$, where $\ominus$ denotes the symmetric difference operator. The union of all such graphs is $G_{S,T} = (V_{S,T}, E_{S,T})$ with $V_{S,T} = \bigcup_{k \in \mathbb{N}} V_{S,T}^k$ and $E_{S,T} = \bigcup_{k \in \mathbb{N}} E_{S,T}^k$. Sufficient facet-defining conditions for the inequalities corresponding to (5) in the Euclidean case are provided in [6] and strongly inspire the following result.

Theorem 9. *Let $(S \subseteq V_B, T \subseteq V_R)$ be linearly inseparable. Then, (5) defines a facet of P_1 if all the following conditions hold:*

(F1) for each $k > 1$, $V^k_{S,T} = \emptyset$ or ($|V^k_{S,T}| > 1$ and $G^k_{S,T}$ is connected), and

(F2) for each $i \in V_B \backslash S$ (resp. $j \in V_R \backslash T$), there exists an $N \in \mathcal{N}^*(S,T)$ such that $(S \cup \{i\} \backslash N, T \backslash N)$ (resp. $(S \backslash N, T \cup \{j\} \backslash N)$) is linearly separable.

Proof. Let F be the face of P_1 defined by (5), and suppose that $\boldsymbol{\lambda}^\top \boldsymbol{y} = \lambda_0$ for every $\boldsymbol{y} \in F$. We show that each entry of $\boldsymbol{\lambda}$ is a multiple of the corresponding coefficient in (5) [20].

Claim 10. *$G^k_{S,T}$ is connected, for every $k \in \mathbb{N}$.*

Proof. $G^1_{S,T}$ is complete or empty and, for $k > 1$, condition (F1) applies. □

There are three cases to analyze. First, let $i, j \in V^k_{S,T}$, $k \in \mathbb{N}$. If $ij \in E^k_{S,T}$, then there are two perfect $\mathcal{N}$-sets N_i and N_j such that $i \in N_i$, $j \in N_j$, and $N_i \ominus N_j = \{i, j\}$. Then, $\boldsymbol{y}^{N_i}, \boldsymbol{y}^{N_j} \in F$ by Proposition 8. Besides, these two solutions only differ in the variables y_i and y_j, and so we have $\lambda_i = \lambda_j$. If $ij \notin E^k_{S,T}$, we still get $\lambda_i = \lambda_j$ since $G^k_{S,T}$ is connected by Claim 10.

Second, let $i, j \in S \cup T$ such that $\nu_i \neq \nu_j$, $\nu_i, \nu_j \in \mathbb{N}$. If i has at least one neighbor in $G^{\nu_i}_{S,T}$, then condition (F1) ensures the existence of a perfect $\mathcal{N}$-set $N_i \subseteq S \cup T$ with $i \in N_i$. On the other hand, if i is an isolated vertex in $G^{\nu_i}_{S,T}$, then condition (F1) yields $\nu_i = 1$, hence $N_i := \{i\}$ is a perfect $\mathcal{N}$-set. The same argument ensures the existence of a perfect $\mathcal{N}$-set $N_j \subseteq S \cup T$ such that $j \in N_j$. Proposition 8 implies that $\boldsymbol{y}^{N_i} \in F$ and $\boldsymbol{y}^{N_j} \in F$. So $\boldsymbol{\lambda}^\top \boldsymbol{y}^{N_i} = \boldsymbol{\lambda}^\top \boldsymbol{y}^{N_j}$, and hence $\nu_i \lambda_i = \nu_j \lambda_j$.

Finally, let $i \in V_B \backslash S$. By condition (F2), let $N \in \mathcal{N}^*(S,T)$ be such that $(S \cup \{i\} \backslash N, T \backslash N)$ is linearly separable. This last condition implies that $\boldsymbol{y}^N - \boldsymbol{e}^i$ is a feasible solution. Since N is perfect, we have $\boldsymbol{y}^N \in F$ by Proposition 8 and $\boldsymbol{y}^N - \boldsymbol{e}^i \in F$ because $i \notin S \cup T$. Hence, we get $\lambda_i = 0$. A similar argument allows us to conclude that $\lambda_j = 0$ for every $j \in V_R \backslash T$. □

3.4 Facet-Defining Minimal $\mathcal{N}$-Set Inequalities

It is known that all inequalities with integer coefficients and righthand side equal to 1 defining facets of P_1 are included in (2)-(3) [4]. As discussed in Subsection 3.1, they are defined by linearly inseparable pairs. From the results of the previous subsection, we can conclude that they are exactly the $\mathcal{N}$-set inequalities given by *minimal* linearly inseparable pairs. Using Theorem 9, we can get necessary and sufficient facetness conditions in this case.

Corollary 11. *An $\mathcal{N}$-set inequality (5) induced by a minimal linearly inseparable pair $(S \subseteq V_B, T \subseteq V_R)$ is facet-defining for P_1 if, and only if,*

(F_E) for each $i \in V_B \backslash S$ (resp. $j \in V_R \backslash T$), there exists an $\ell \in S \cup T$ such that $(S \cup \{i\} \backslash \{\ell\}, T \backslash \{\ell\})$ (resp. $(S \setminus \{\ell\}, T \cup \{j\} \setminus \{\ell\})$) is linearly separable.

Proof. First, assume that (F_E) is satisfied. Since (S,T) is minimal, $V^k_{S,T} = \emptyset$ for all $k > 1$ and so (F1) holds. In addition, (F_E) directly implies (F2). Thus, the $\mathcal{N}$-set inequality for (S,T) is facet-defining.

Conversely, suppose that (F_E) is violated by (S,T). Let $i \in V_B \setminus S$ be such that $(S' \setminus \{\ell\}, T \setminus \{\ell\})$, for all $\ell \in S \cup T$, is linearly inseparable, where $S' = S \cup \{i\}$. It turns out that (S',T) is linearly inseparable. Besides, for all $j \in S' \cup T$,

$$\nu'_j := \min\{|N| \mid N \in \mathcal{N}(S',T), j \in N\} = 2$$

since (S,T) is linearly inseparable and $\{i,j\} \in \mathcal{N}(S',T)$ by the minimality of (S,T). Consequently,

$$\sum_{j \in S' \cup T} y_j \geq 2$$

is valid for P_1. This inequality together with $y_i \geq 1$ dominate the $\mathcal{N}$-set inequality for (S,T), which then does not define a facet of P_1. □

Two special cases of minimal $\mathcal{N}$-set inequalities that can be separated in polynomial time are illustrated in Figure 3. The structure in Figure 3a is a special case of (2) when $|S| = 2$ and $i \in H[S]$ is replaced by the stronger condition $i \in D[S]$. We refer to this case as *generalized* 3*-path inequality* which, in general terms, is written as

$$y_j + y_{j'} + y_i \geq 1, \qquad i \in V_{BR}, \{j,j'\} \subseteq V_{\bar{K}(i)} \text{ such that } i \in D(j,j'). \tag{6}$$

Similarly, the structure in Figure 3b produces an inequality of type (3) with $|S| = |T| = 2$ that satisfies stronger conditions (with respect to $H[S] \cap H[T] \cap V_N \neq \emptyset$). Precisely, it can be written as

$$\begin{aligned} y_v + y_{v'} + y_w + y_{w'} \geq 1, \quad & \{v,v'\} \subseteq V_B, \{w,w'\} \subseteq V_R \text{ such that } D[v,v'] \cap D[w,w'] \cap V_N \neq \emptyset \text{ and} \\ & D[v,v'] \cap \{w,w'\} = D[w,w'] \cap \{v,v'\} = \emptyset. \end{aligned} \tag{7}$$

It is called *X-swing inequality*.

3.5 Valid and Facet-Defining Non-Minimal $\mathcal{N}$-Set Inequalities

In remainder of the section, we analyze some non-minimal $\mathcal{N}$-set inequalities, *i.e.* $\mathcal{N}$-set inequalities different from those of the integer formulation. The first one is the rank-1 Chvátal-Gomory inequality depicted in Figure 3c. It is the counterpart of a facet-defining inequality for the formulation described in [1].

Proposition 12. *If $v, v' \in V_B$ and $w, w' \in V_R$ are distinct vertices such that $\{v,v'\} \subseteq H[w,w']$ and $\{w,w'\} \subseteq H[v,v']$, then* the generalized C_4 inequality

$$y_v + y_{v'} + y_w + y_{w'} \geq 2 \tag{8}$$

is an $\mathcal{N}$-set inequality and facet-defining for P_1.

Proof. Let $(S = \{v,v'\}, T = \{w,w'\})$. The fact that any size-2 subset of $S \cup T$ defines a perfect $\mathcal{N}$-set for (S,T) has several consequences. First, $\nu_i = 2$ for every $i \in S \cup T$, which means that (8) is an $\mathcal{N}$-set inequality (recall that Proposition 7 ensures validity). Second, the safe graph $G^2_{S,T}$

is complete, $|V^2_{S,T}| = 4$, and $V^k_{S,T} = \emptyset$ for all $k > 2$. Thus, Condition (F1) of Theorem 9 holds for (8). Finally, T is a perfect $\mathcal{N}$-set such that $(S \cup \{i\}, \emptyset)$, for all $i \in V_B \setminus S$, is linearly separable, and S is a perfect $\mathcal{N}$-set such that $(\emptyset, T \cup \{j\})$, for all $j \in V_R \setminus T$, is linearly separable. Therefore, Condition (F2) of Theorem 9 holds. It follows that (8) is facet-defining. □

Two final remarks in connection with the generalized C_4 inequality are worthwhile. First, note that it results from a configuration of geodesic convex combinations that cannot occur in the Euclidean space. Note in addition that it is a $\{0, \frac{1}{3}\}$-Chvátal-Gomory cut with multiplier $1/3$ for each $\mathcal{N}$-set inequality defined for any 3-path in $\{v, w, v', w'\}$.

$\mathcal{N}$-set inequalities induced by structures depicted in Figure 3a and Figure 3d are discussed below.

Proposition 13. *Let $K \in \{B, R\}$, $S \subseteq V_K$, $|S| \geq 2$, and $j \in V_{\bar{K}}$ be such that $j \in H[i, i']$ for all $i, i' \in S$, $i \neq i'$. Then,* the star tree inequality

$$\sum_{i \in S} y_i + (|S| - 1) y_j \geq (|S| - 1) \tag{9}$$

is an $\mathcal{N}$-set inequality and is valid for P_1. Moreover, if $j' \notin H[S]$ or $S \setminus H[j, j'] \neq \emptyset$, for every $j' \in V_{\bar{K}} \setminus \{j\}$, then (9) defines a facet of P_1.

Proof. By definition, the pair $(S, T = \{j\})$ is linearly inseparable and $|S| \geq 2$. For this pair, it turns out that $\nu_j = 1$ and $\nu_i = \nu = |S| - 1$ for all $i \in S$. Thus, by Proposition 7, inequality (9) is an $\mathcal{N}$-set inequality valid for P_1. In addition, $G^{\nu}_{S,T}$ is a complete safe graph since any size-ν subset of S is a perfect $\mathcal{N}$-set. Thus, we can conclude that Condition (F1) of Theorem 9 holds for (9). Finally, consider Condition (F2) of Theorem 9. It trivially holds for any vertex in $V_K \setminus S$ since $\nu_j = 1$. Now, let $j' \in V_{\bar{K}} \setminus \{j\}$. By the facetness hypothesis, $(S, \{j'\})$ or $(i, \{j, j'\})$ for some $i \in S$ is linearly separable. In both cases, Condition (F2) is satisfied. Therefore, (9) defines a facet of P_1. □

The structure considered in the following result is illustrated in Figure 3e.

Proposition 14. *Let $K \in \{B, R\}$, $(S \subseteq V_K, T \subseteq V_{\bar{K}})$ be a linearly inseparable pair, $|S|, |T| \geq 2$, $| \, |S| - |T| \, | = 1$, and $S = S_1 \cup S_2$ and $T = T_1 \cup T_2$ be two partitions such that $|S_1| - |S_2| = 1$ and $|T_1| = |T_2|$. Moreover, assume that the following properties hold:*

(a) there exist vertices $m_1, m_2 \in V_N$ (not necessarily distinct) where $m_\ell \in H[v, v']$ for all $\ell \in \{1, 2\}$ and all v, v' such that $v, v' \in S_\ell$ or $v, v' \in T_\ell$, and

(b) $\{m_1, m_2\}$ is contained in $H[v, v']$ for all $(v, v') \in S_1 \times S_2$ and $(v, v') \in T_1 \times T_2$.

Then,

$$(|T| - 1) \sum_{i \in S} y_i + (|S| - 1) \sum_{j \in T} y_j \geq (|S| - 1)(|T| - 1) \tag{10}$$

is valid for P_1. Moreover, if

1. *$|S \setminus H[T]| \geq 2$, $|T \setminus H[S]| \geq 2$,*

2. *for every* $j' \in V_{\bar{K}} \setminus T$, *there exists* $j \in T$ *such that* $S \cap H[T \ominus \{j, j'\}] \neq \emptyset$ *or* $H[S] \cap H[j, j'] = \emptyset$, *and*

3. *for every* $i' \in V_K \setminus S$, *there exists* $i \in S$ *such that* $H[S \ominus \{i, i'\}] \cap T \neq \emptyset$ *or* $H[i, i'] \cap H[T] = \emptyset$,

then (10) *defines a facet of* P_1.

Proof. From property (b), we first observe that $(\{v, v'\}, \{w, w'\})$ is an inseparable pair for all $v \in S_1$, $v' \in S_2$, $w \in T_1$, and $w' \in T_2$. It follows that either S_1, S_2, T_1, or T_2 is contained in any $\mathcal{N}$-set N. Among these sets, if N contains only S_ℓ, for some $\ell \in \{1, 2\}$, then it must contain at least $|S| - 1$ elements from S, otherwise two elements in the other part of S together with one vertex from T_1 and one vertex from T_2 imply that N is not a separator due to properties (a) and (b). Similarly, we have that, if N contains only T_ℓ, for some $\ell \in \{1, 2\}$, then it must contain at least $|T| - 1$ elements from T. Recall that $|S| \geq |T| - 1$ and $|T| \geq |S| - 1$ because $||S| - |T|| = 1$.

To show that $\nu_i \geq |S| - 1$ for each $i \in S$, consider an $\mathcal{N}$-set N containing i. We have to prove that $|N| \geq |S| - 1$. If N contains only one of the sets S_1, S_2, T_1 and T_2, the above results imply that $|N| \geq |S| - 1$, since $i \notin T$. If $S_\ell \cup T_{\ell'} \subseteq N$ for some $\ell, \ell' \in \{1, 2\}$, then $|N| \geq |S_\ell| + |T_{\ell'}| \geq (|S| - 1)/2 + |T|/2 \geq |S| - 1$. If $S_1 \cup S_2 \subseteq N$, then $|N| = |S|$. If $T_1 \cup T_2 \subseteq N$, then $|N| = |T| \geq |S| - 1$. In any case, we conclude that $\nu_i \geq |S| - 1$ for each $i \in S$. Similarly, we get $\nu_j \geq |T| - 1$ for all $j \in T$. Therefore, by Proposition 7, the inequality $\frac{1}{|S|-1} \sum_{i \in S} y_i + \frac{1}{|T|-1} \sum_{j \in T} \geq 1$, or equivalently (10), is valid.

To show facetness, we first prove that $\nu_i = |S| - 1$ and $\nu_j = |T| - 1$, for each $i \in S$ and $j \in T$. Indeed, condition 1) imply that any size-$(|S|-1)$ subset of S and any size-$(|T|-1)$ subset of T is a perfect $\mathcal{N}$-set. Then, we can apply Theorem 9. Besides, these $\mathcal{N}$-sets show that Condition (F1) is satisfied. Now, consider $h \in V_{BR} \setminus (S \cup T)$. By conditions 2) and 3), either $(S, \{h, j\})$ or $(S \cup \{h\}, \{j\})$ for some $j \in T$, or $(\{h, i\}, T)$ or $(\{i\}, T \cup \{h\})$ for some $i \in S$, is linearly separable. Then, Condition (F2) of Theorem 9 is satisfied, and so (10) is facet-defining. □

As a last special case, we show next an $\mathcal{N}$-set inequality resulting from the structure illustrated in Figure 4, called *alternating path inequality*.

Proposition 15. *Let* $S = \{i_1, \ldots, i_{\ell+1}\} \subseteq V_K$, $T = \{j_1, \ldots, j_\ell\} \subseteq V_{\bar{K}}$, $\ell \geq 1$ *and* $K \in \{B, R\}$, *such that* $\langle i_1, j_1, \ldots, i_\ell, j_\ell, i_{\ell+1} \rangle$ *is a sequence contained in a geodesic between* i_1 *and* $i_{\ell+1}$ *in* G. *The* alternating path *inequality*

$$\sum_{i \in S \cup T} y_i \geq \ell \tag{11}$$

is valid for P_1. *Moreover, it is facet-defining if Condition (F2) of Theorem 9 holds as well as the following conditions:*

1. $i_k \notin H[j_1, \ldots, j_{k-1}] \cup H[j_k, \ldots, j_\ell]$, $1 \leq k \leq \ell + 1$, *and*

2. $j_k \notin H[i_1, \ldots, i_k] \cup H[i_{k+1}, \ldots, i_{\ell+1}]$, $1 \leq k \leq \ell$.

Proof. To prove validity, we use induction on $|T| = |S| - 1$. When $|T| = 1$, the corresponding inequalities are exactly the ones in (6), which are already in the formulation, and thus are valid.

Assume validity for $|T| \leq \ell - 1$, $\ell \geq 2$. Now, consider $|T| = \ell$, i.e. the alternating path inequality defined by a sequence $\langle i_1, j_1, \ldots, i_\ell, j_\ell, i_{\ell+1} \rangle$. By the induction hypothesis, the inequalities defined by the following subsequences are valid:

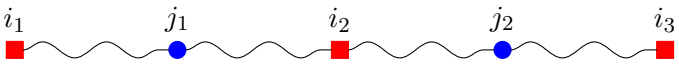
- $\langle i_1, j_1, \ldots, j_{\ell-1}, i_\ell \rangle$: $\sum_{k=1}^{\ell} y_{i_k} + \sum_{k=1}^{\ell-1} y_{j_k} \geq \ell - 1$;
- $\langle i_2, j_2, \ldots, j_\ell, i_{\ell+1} \rangle$: $\sum_{k=2}^{\ell+1} y_{i_k} + \sum_{k=2}^{\ell} y_{j_k} \geq \ell - 1$;
- $\langle i_1, j_\ell, i_{\ell+1} \rangle$: $y_{i_1} + y_{j_\ell} + y_{i_{\ell+1}} \geq 1$;

By summing up these three valid inequalities and $y_{j_1} \geq 0$, then dividing by 2 the resulting inequality and rounding up the right-hand side, we get exactly (11). Therefore, it is valid.

To prove facetness, we first show that condition 1 ensures that (11) is an $\mathcal{N}$-set inequality. As (S, T) is linearly inseparable, it suffices to prove that $\nu_u = \ell$, for every $u \in S \cup T$. Since the vertices in the sequence alternate colors and $|S| > |T| = \ell$, it holds that $(S \setminus N, T \setminus N)$ is still linearly inseparable for any set N with less than ℓ vertices. Then, $\nu_u \geq \ell$, for every $u \in S \cup T$. On the other hand, for every $k \in \{1, \ldots, \ell\}$, observe that T is an $\mathcal{N}$-set that contains j_k, thus implying $\nu_{j_k} \leq \ell$, and so $\nu_{j_k} = \ell$. Similarly, both $S \setminus \{i_1\}$ and $S \setminus \{i_{\ell+1}\}$ are $\mathcal{N}$-sets, since conditions 1 holds. So, for every $k \in \{1, \ldots, \ell + 1\}$, there is an $\mathcal{N}$-set N with $|N| = \ell$ that contains i_k, which implies that $\nu_{i_k} = \ell$.

Now, it remains to show that Condition (F1) of Theorem 9 is satisfied by (S, T), or equivalently that $G^\ell_{S,T} = (V^\ell_{S,T}, E^\ell_{S,T})$ is connected. As already shown, $V^\ell_{S,T} = S \cup T$. Thus, it suffices to prove that $i_k j_k$ and $j_k i_{k+1}$ belong to $E^\ell_{S,T}$, for all $k \in \{1, \ldots, \ell\}$. By conditions 1 and 2, for all $k \in \{1, \ldots, \ell+1\}$, we have that $\{j_1, \ldots, j_{k-1}, i_k, \ldots, i_{\ell+1}\}$ and $\{i_1, \ldots, i_k, j_k, \ldots, j_\ell\}$ are linearly separable. Therefore, $N_k := \{i_1, \ldots, i_{k-1}, j_k, \ldots, j_\ell\}$ and $N'_k := \{j_1, \ldots, j_{k-1}, i_{k+1}, \ldots, i_{\ell+1}\}$ belong to $\mathcal{N}^*(S, T)$, for all $k \in \{1, \ldots, \ell + 1\}$. Besides, since $N_{k+1} \ominus N_k = \{i_k, j_k\}$ and $N'_{k+1} \ominus N'_k = \{j_k, i_{k+1}\}$, we conclude that $i_k j_k, j_k i_{k+1} \in E^\ell_{S,T}$, for all $k \in \{1, \ldots, \ell\}$. □

Figure 4 – Alternating path resulting in the valid inequality $\sum_{t=1}^{2}(y_{i_t} + y_{j_t}) + y_{i_3} \geq 2$ (Proposition 15). Snaked segments represent shortest paths of their endpoints.



4 A Compact Formulation for the 2-GC Problem

4.1 Integer Formulation

The second integer linear formulation is obtained by including additional variables so as to reduce the number of constraints to a polynomial order. The new binary variables, z, are used to determine if a vertex belongs to the convex hull of the non-outliers of a given class. More precisely, in a 2-GC feasible solution (A_B, A_R), for each $K \in \{B, R\}$ and $i \in V$, we define a binary variable $z_{K,i} = 1$ which is set to 1 if $i \in H[A_K]$. Thus, the feasible solutions of the new formulation are defined by the binary vectors $\boldsymbol{y} \in \mathbb{B}^n$ and $\boldsymbol{z} \in \mathbb{B}^{2|V|}$ such that

$$y_i \geq z_{\bar{K},i}, \qquad i \in V_K, K \in \{B, R\} \tag{12}$$

$$y_i + z_{K,i} \geq 1, \qquad i \in V_K, K \in \{B, R\} \tag{13}$$

$$z_{B,i} + z_{R,i} \leq 1, \qquad i \in V_N \tag{14}$$

$$z_{K,h} + z_{K,j} - z_{K,i} \leq 1, \qquad K \in \{B, R\}, h, i, j \in V : i \in D(h, j) \tag{15}$$

This formulation is denoted ILP2. Next, we show that these inequalities model conditions (C1)–(C3) by showing how (2)–(3) and (12)–(15) are related.

Proposition 16. *Let* $F_1 = \{\boldsymbol{y} \in \mathbb{B}^n \mid (2),\ (3)\}$ *and* $F_2 = \{\boldsymbol{y} \in \mathbb{B}^n, \boldsymbol{z} \in \mathbb{B}^{2|V|} \mid (12)\text{–}(15)\}$. *Then,* $F_1 = \text{proj}_{\boldsymbol{y}}(F_2)$.

Proof. Let $\boldsymbol{y} \in F_1$ and $A_K = \{i \in V_K : y_i = 0\}$, for $K \in \{B, R\}$. By Proposition 2, A_B and A_R satisfy conditions (C1)–(C3). We claim that $(\boldsymbol{y}, \boldsymbol{z}) \in F_2$, where $\boldsymbol{z} \in \mathbb{B}^{2|V|}$ is such that, for all $K \in \{B, R\}$ and $i \in V$, $z_{K,i} = 1$ if, and only if, $i \in H[A_K]$. Let $i \in V_K$, $K \in \{B, R\}$. If $y_i = 1$, constraints (12) and (13) are trivially satisfied. Otherwise, $i \in A_k$ which gives $z_{K,i} = 1$. Besides, condition (C1) or (C2) implies $i \notin H[A_{\bar{K}}]$, and so $z_{\bar{K},i} = 0$. Again, constraints (12) and (13) are satisfied by $(\boldsymbol{y}, \boldsymbol{z})$. Inequality (14) stems directly from Condition (C3). Finally, by definition, $D(h, j) \subseteq H[h, j]$, which means that if $z_{K,h} = z_{K,j} = 1$ then $z_{K,i} = 1$, for all $h, i, j \in V$ such that $i \in D(h, j)$. This shows that constraints (15) are satisfied. Therefore, $F_1 \subseteq \text{proj}_{\boldsymbol{y}}(F_2)$.

Conversely, we claim that $(\boldsymbol{y}, \boldsymbol{z}) \in F_2$ yields that $\boldsymbol{y}$ satisfies (2)-(3). First, let $i \in V_K$ and $S \subseteq V_{\bar{K}}$ such that $i \in H[S]$. Note that $|S| \geq 2$. If $i' \in S$ is such that $y_{i'} = 0$, then $z_{\bar{K},i'} = 1$ by (13). Then, since $i \in H[S]$, we can use (15) to conclude that $z_{\bar{K},i} = 1$. So, (12) ensures that $y_i = 1$, showing that (2) is satisfied. Now, let $S \subseteq V_B$ and $T \subseteq V_R$ such that $H[S] \cap H[T] \cap V_N \neq \emptyset$. Note that $|S| \geq 2$ and $|T| \geq 2$. Suppose by contradiction that (3) is violated, i.e. $y_i = 0$ for all $i \in S \cup T$. By (13), it follows that $z_{B,i} = 1$ for all $i \in S$, and $z_{R,j} = 1$ for all $j \in T$. If $i \in H[S] \cap H[T] \cap V_N$, then (15) implies that $z_{R,i} = z_{B,i} = 1$, which contradicts (14). Therefore, we conclude that $\boldsymbol{y} \in F_2$. □

Propositions 2 and 16 imply the correctness of (12)–(15). Although the relation established in Proposition 16, there are examples where the linear relaxation of (2)–(3) provides a better bound than (12)–(15), and vice-versa.

4.2 Polyhedra

We discuss next some structural properties of $P_2 = \text{conv}\left\{(\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{B}^n \times \mathbb{B}^{2n_z} \mid (12)\text{–}(15)\right\}$, where $n_z = |V|$. Here, we adopt the same notation of $\mathbf{0}$, $\mathbf{1}$, $\boldsymbol{e}^i$, and $\bar{\boldsymbol{e}}^i$ defined in Section 3.2, with the addition of $\boldsymbol{e}^{K,i}$ to represent the unit vector with $z_{K,i} = 1$ and $\bar{\boldsymbol{e}}^{K,i} = \mathbf{1} - \boldsymbol{e}^{K,i}$.

Proposition 17. P_2 *is full-dimensional.*

Proof. Consider the following $n + 2n_z + 1$ points in P_2: $\boldsymbol{v}^{K,i} = (\mathbf{1}, \boldsymbol{e}^{K,i})$, for all $i \in V$ and $K \in \{B, R\}$, $\boldsymbol{w}^0 = (\mathbf{1}, \mathbf{0})$, and $\boldsymbol{w}^i = (\bar{\boldsymbol{e}}^i, \boldsymbol{e}^{K,i})$, for all $(K = B, i \in V_B)$ and $(K = R, i \in V_R)$. Suppose that they all satisfy the equality $\boldsymbol{\pi}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} = \pi_0$, for some $(\boldsymbol{\pi}, \boldsymbol{\mu}, \pi_0) \in \mathbb{R}^n \times \mathbb{R}^{2n_z} \times \mathbb{R}$. Let $i \in V$ and $K \in \{B, R\}$. By subtracting the equalities for $\boldsymbol{v}^{K,i}$ and $\boldsymbol{w}^0$, we get $\boldsymbol{\mu}^\top \boldsymbol{e}^{K,i} = 0$, that is, $\mu_{K,i} = 0$. Similarly, if $i \in V_{BR}$, then the difference between the equalities for $\boldsymbol{v}^{K,i}$ and $\boldsymbol{w}^i$ gives $\pi_i = 0$. Therefore, $\boldsymbol{\pi} = \mathbf{0}$, $\boldsymbol{\mu} = \mathbf{0}$, $\pi_0 = 0$, and the points are affinely independent. □

By Proposition 16, any valid inequality for P_1, in special the ones described in Subsection 3, is also valid for P_2. However, the facetness conditions for P_1 are not directly transferred to P_2, even for the bounding constraints. In the remaining of this section, we show facets of P_2 defined by the constraints of ILP2. The affinely independent points defined in the proof of Proposition 17 imply the following.

Proposition 18. *The following inequalities are facet-defining for P_2:*

1. *$y_i \leq 1$, for all $i \in V_{BR}$, and*

2. *$z_{K,i} \geq 0$, for all $K \in \{B, R\}$ and $i \in V_N \cup V_{\bar{K}}$.*

To show other facet-defining bounding inequalities, we need the following definition. It will also be useful in the next section. Let $\mathcal{U}$ be a set of subsets of V and $V' = V \setminus \bigcup_{U \in \mathcal{U}} U$. The graph G is said to be $\mathcal{U}$-arrangeable if the vertices in V' can be ordered as $i_1, \ldots, i_p$, $p = |V'|$, such that, for all $k \in \{1, \ldots, p\}$, there exists $U \in \mathcal{U}$ satisfying $H[U \cup \{i_k\}] \subseteq U \cup \{i_1, \ldots, i_k\}$.

Proposition 19. *For every $K \in \{B, R\}$ and $i \in V_K$, $z_{K,i} \leq 1$ is facet-defining for P_2 if and only if G is $\{i\}$-arrangeable.*

Proof. Let $K \in \{B, R\}$, $i \in V_K$, and $F = \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 : z_{K,i} = 1\}$. Assume that $F \subseteq F' := \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 \mid \boldsymbol{\pi}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} = \pi_0\}$, for some $(\boldsymbol{\pi}, \boldsymbol{\mu}, \pi_0) \in \mathbb{R}^n \times \mathbb{R}^{2n_z} \times \mathbb{R}$. It is sufficient to prove that $\mu_{\bar{K},j} = 0$ for all $j \in V$, $\mu_{K,j} = 0$ for all $j \in V \setminus \{i\}$, and $\boldsymbol{\pi} = \mathbf{0}$. We consider the following cases (in each of them we present two points in $F \subseteq F'$ to get the desired result):

$\mu_{K,j} = 0$ **for all** $j \in V \setminus \{i\}$**:** By hypothesis, there is an ordering $i_1, \ldots, i_p$ of $V \setminus \{i\}$ such that $H[i, i_k] \subseteq \{i, i_1, \ldots, i_k\}$, for all $k \in \{1, \ldots, p\}$. Then, using induction on k, we conclude that $\mu_{K,i_k} = 0$ for all $k \in \{i, \ldots, p\}$, since $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\mathbf{1}, \sum_{j \in H[i,i_k]} \boldsymbol{e}^{K,j}) \in F$ and $H[i, i_k] \setminus \{i\} \subseteq \{i_1, \ldots, i_k\}$.

$\mu_{\bar{K},j} = 0$ **for all** $j \in V$**:** the points $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\mathbf{1}, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},j}) \in F$ imply that $\boldsymbol{\mu}^\top \boldsymbol{e}^{\bar{K},j} = \mu_{\bar{K},j} = 0$. Note that the second point belongs to F even if $j = i$;

$\pi_j = 0$ **for all** $j \in V_{BR}$**:** If $j \in V_{\bar{K}}$, points $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\bar{\boldsymbol{e}}^j, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},j}) \in F$ lead to $\pi_j = 0$. If $j \in V_K$, points $(\mathbf{1}, \sum_{h \in H[i,j]} \boldsymbol{e}^{K,h}), (\bar{\boldsymbol{e}}^j, \sum_{h \in H[i,j]} \boldsymbol{e}^{K,h}) \in F$ show that $\pi_j = 0$.

Now, assume that G is not $\{i\}$-arrangeable. We claim that there are distinct vertices $j, j' \in V \setminus \{i\}$ such that $j' \in H[i, j]$ and $j \in H[i, j']$. Therefore, $z_{K,i} + z_{K,j} - z_{K,j'} \leq 1$ and $z_{K,i} + z_{K,j'} - z_{K,j} \leq 1$ are valid for P_2. Their sum is $2z_{K,i} \leq 2$, which shows that $z_{K,i} \leq 1$ does not define a facet in this case.

We prove the claim by contrapositive. Suppose that, for all distinct vertices $j, j' \in V \setminus \{i\}$, $j' \notin H[i, j]$ or $j \notin H[i, j']$. Define a binary relation $\leq$ on $V \setminus \{i\}$ such that $j \leq j'$ if and only if $j \in H[i, j']$. We show that $\leq$ is a strict partial order. Besides being trivially reflexive, it is antisymmetric due to the hypothesis. Moreover, it is transitive because $j' \in H[i, j'']$ and $j \in H[i, j']$ imply $j \in H[i, j'']$, and so $j \leq j' \leq j''$ implies that $j \leq j''$. Now, let $i_1, \ldots, i_{|V|-1}$ be an ordering of $V \setminus \{i\}$ which is an extension of $\leq$. Let $k \in \{1, \ldots, |V| - 1\}$. For any $i_\ell \in H[i, i_k]$, the order definition ensures that $i_\ell \leq i_k$. Therefore, $H[i, i_k] \subseteq \{i, i_1, \ldots, i_k\}$, and so G is $\{i\}$-arrangeable. □

In the sequel, we analyze constraints (12)-(14). Generalizations of constraints (15) will be derived in the next section (see Propositions 23 and 24).

Proposition 20. *For all $i \in V_N$, $z_{B,i} + z_{R,i} \leq 1$ is facet-defining for P_2.*

Proof. For $K \in \{B, R\}$, consider the $n + 2n_z$ points $(\mathbf{1}, \boldsymbol{e}^{K,i})$, $(\mathbf{1}, \boldsymbol{e}^{\bar{K},i} + \boldsymbol{e}^{K,j})$ for all $j \in V \backslash \{i\}$, and $(\bar{\boldsymbol{e}}^j, \boldsymbol{e}^{\bar{K},i} + \boldsymbol{e}^{K,j})$ for all $j \in V_K \backslash \{i\}$ are feasible solutions that satisfy $z_{B,i} + z_{R,i} \leq 1$ at equality. Besides, they are affinely independent. □

Proposition 21. *For all $i \in V_K$ and $K \in \{B, R\}$, $y_i \geq z_{\bar{K},i}$ and $y_i + z_{K,i} \geq 1$ are facet-defining for P_2.*

Proof. The $n + 2n_z$ affinely independent points of P_2 satisfying $y_i \geq z_{\bar{K},i}$ at equality are as follows: $(\bar{\boldsymbol{e}}^i, \boldsymbol{e}^{K,i})$, $(\mathbf{1}, \boldsymbol{e}^{\bar{K},i})$, $(\mathbf{1}, \boldsymbol{e}^{\bar{K},i} + \boldsymbol{e}^{K,j})$ for all $j \in V$, $(\bar{\boldsymbol{e}}^i, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},j})$ for all $j \in V \setminus \{i\}$, $(\bar{\boldsymbol{e}}^j, \boldsymbol{e}^{\bar{K},i} + \boldsymbol{e}^{K,j})$ for all $j \in V_K \setminus \{i\}$, and $(\bar{\boldsymbol{e}}^j - \boldsymbol{e}^i, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},j})$ for all $j \in V_{\bar{K}}$.

For $y_i + z_{K,i} \geq 1$, the $n + 2n_z$ affinely independent points are $(\mathbf{1}, 0)$, $(\mathbf{1}, \boldsymbol{e}^{K',j})$ for all $j \in V$ and $K' \in \{B, R\}$ such that $(K', j) \neq (K, i)$, and $(\bar{e}^j, e^{K(j),j})$ for all $j \in V_{BR}$. □

We close this subsection by indicating bounding inequalities that do not define facets of P_2.

Proposition 22. *For all $K \in \{B, R\}$, the constraints below do not define facets of P_2:*

1. *$y_i \geq 0$, $z_{\bar{K},i} \leq 1$, and $z_{K,i} \geq 0$, for all $i \in V_K$, and*
2. *$z_{K,i} \leq 1$, for all $i \in V_N$.*

Proof. By (12), $y_i = 0$ implies $z_{\bar{K},i} = 0$ (and $z_{\bar{K},i} = 1$ implies $y_i = 1$), for all $i \in V_K$. So, $y_i \geq 0$ and $z_{\bar{K},i} \leq 1$ cannot define facet of P_2. By (13), since $z_{K,i} = 0$ implies $y_i = 1$, $z_{K,i} \geq 0$ does not define a facet of P_2. Finally, $z_{K,i} \leq 1$ is dominated by $z_{K,i} + z_{\bar{K},i} \leq 1$, for all $i \in V_N$. □

4.3 Lifting N-set Inequalities from P_1

As already mentioned, any inequality $\boldsymbol{\pi}^\top \boldsymbol{y} \geq \pi_0$ valid for P_1 is also valid for P_2. As it may not induce a facet of P_2 even if it does for P_1, we could think of a lifting strategy. Due to (12) and (13), a straightforward strategy when $\pi_i \geq 0$ would be to replace y_i by $z_{\bar{K},i}$, if $i \in V_{\bar{K}}$, or by $1 - z_{K,i}$, if $i \in V_K$. We show here some of the facet-defining N-set inequalities that still remain valid for P_2 after such a lifting. We also show sufficient conditions to get facets of P_2 from them.

We start by resuming the structure associated with star tree inequality (9) illustrated in Figure 3a and Figure 3d.

Proposition 23. *Let $K \in \{B, R\}$, $S \subseteq V$, and $j \in V \setminus S$ be such that $j \in H[i, i']$ for all $i, i' \in S$, $i \neq i'$. Then,*

$$\sum_{i \in S} z_{K,i} - (|S| - 1) z_{K,j} \leq 1 \tag{16}$$

is valid for P_2. Moreover, for $S \subseteq V_K$ and $j \in V_{\bar{K}}$, (16) dominates (9). Also, let $\mathcal{U} = \{\{i\} : i \in S\} \cup \{S \cup \{j\}\}$. If G is $\mathcal{U}$-arrangeable, then (16) is facet-defining for P_2.

Proof. Let $(\boldsymbol{y}, \boldsymbol{z})$ be a feasible solution. If $\sum_{i \in S} z_{K,i} \leq 1$, then inequality (16) trivially holds. So, suppose that $\sum_{i \in S} z_{K,i} > 1$. Then, there exist $i, i' \in S$, $i \neq i'$, such that $z_{K,i} = z_{K,i'} = 1$. By hypothesis, $j \in H[i, i']$, and by (15), $z_{K,j} = 1$. Therefore, $(|S|-1) z_{K,j} = |S| - 1 \geq \sum_{i \in S} z_{K,i} - 1$, and so (16) holds.

Now consider (16) for $S \subseteq V_K$ and $j \in V_{\bar{K}}$. Then, using (12) and (13), we get

$$1 \geq \sum_{i \in S} z_{K,i} - (|S|-1) z_{K,j} \geq \sum_{i \in S} (1 - y_i) - (|S|-1) y_j,$$

which shows that (16) dominates (9).

To conclude the proof, assume that G is $\mathcal{U}$-arrangeable. Let $F = \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 \mid \sum_{i \in S} z_{K,i} - (|S|-1) z_{K,j} = 1\}$ and assume that $F \subseteq F' := \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 \mid \boldsymbol{\pi}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} = \pi_0\}$, for some $(\boldsymbol{\pi}, \boldsymbol{\mu}, \pi_0) \in \mathbb{R}^n \times \mathbb{R}^{2n_z} \times \mathbb{R}$. Observe that $H[U] \cap (S \cup \{j\}) = U$ and so $(\mathbf{1}, \sum_{i' \in H[U]} \boldsymbol{e}^{K,i'}) \in F$, for every $U \in \mathcal{U}$. The cases analyzed below show that $\boldsymbol{\pi}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} = \pi_0$ is a multiple of (16) at equality.

$\mu_{K,i'} = 0$ **for all** $i' \in V' := V \setminus (S \cup \{j\})$**:** Since G is $\mathcal{U}$-arrangeable, there is an ordering $i_1, \ldots, i_p$ of V', $p = |V'|$, such that, for all $k \in \{1, \ldots, p\}$, there exists $U_k \in \mathcal{U}$ satisfying $H[U_k \cup \{i_k\}] \subseteq U_k \cup \{i_1, \ldots, i_k\}$. Then, $H[U_k] \cap (S \cup \{j\}) = U_k$ and $H[U_k \cup \{i_k\}] \cap (S \cup \{j\}) = U_k$, implying that the points $(\mathbf{1}, \sum_{i' \in H[U_k]} \boldsymbol{e}^{K,i'})$ and $(\mathbf{1}, \sum_{i' \in H[U_k \cup \{i_k\}]} \boldsymbol{e}^{K,i'})$ belong to F. Therefore, as $H[U_k \cup \{i_k\}]$ is the union of $H[U_k]$ with a subset of $\{i_1, \ldots, i_k\}$, we can use induction on k to prove that $\mu_{K,i_k} = 0$ for all $k \in \{1, \ldots, p\}$.

$\mu_{\bar{K},i'} = 0$ **for all** $i' \in V$**:** points $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\mathbf{1}, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},i'}) \in F$, for an arbitrary $i \in S$, imply that $\boldsymbol{\mu}^\top \boldsymbol{e}^{\bar{K},i'} = \mu_{\bar{K},i'} = 0$. Note that the second point belongs to F even if $i' \in S \cup \{j\}$.

$\pi_{i'} = 0$ **for all** $i' \in V_{BR}$**:** If $i' \in V_{\bar{K}}$, the fact that $\mu_{\bar{K},i'} = 0$ and the points $(\mathbf{1}, \boldsymbol{e}^{K,i})$ and $(\bar{\boldsymbol{e}}^{i'}, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},i'})$ belong to F, for any $i \in S$, is sufficient to prove that $\pi_{i'} = 0$. If $i' \in V_K$, then the points $(\mathbf{1}, \sum_{i'' \in H[S \cup \{j,i'\}]} \boldsymbol{e}^{K,i''})$ and $(\bar{e}^{i'}, \sum_{i'' \in H[S \cup \{j,i'\}]} \boldsymbol{e}^{K,i''})$, both in F, show that $\pi_{i'} = 0$.

$\mu_{K,i} = \mu_{K,i'}$ **for all** $i, i' \in S$**,** $i \neq i'$**:** stems from $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\mathbf{1}, \boldsymbol{e}^{K,i'}) \in F$.

$\mu_{K,j} = -(|S|-1)\mu_{K,i}$ **for all** $i \in S$**:** the points $(\mathbf{1}, \boldsymbol{e}^{K,i}), (\mathbf{1}, \sum_{i' \in H[S \cup \{j\}]} \boldsymbol{e}^{K,i'}) \in F$ imply that $\mu_{K,j} + \sum_{i' \in S \setminus \{i\}} \mu_{K,i'} + \sum_{i' \in H[S \cup \{j\}] \setminus (S \cup \{j\})} \mu_{K,i'} = 0$. The claimed result is due to $\mu_{K,i'} = 0$ for all $i' \in V \setminus (S \cup \{j\})$ and $\mu_{K,i} = \mu_{K,i'}$ for all $i, i' \in S$.

□

In [10], a generalization of the convexity constraints (15) was presented for the *Path Convex Recoloring*. As stated below, such generalized inequalities are also valid for P_2 as counterparts of the alternating path inequalities (11).

Proposition 24. *Let $S = \{u_1, v_1, \ldots, u_t, v_t, u_{\ell+1}\}$ such that the sequence $\langle u_1, v_1, \ldots, u_t, v_t, u_{\ell+1} \rangle$, $\ell \geq 1$, is contained in a geodesic between u_1 and $u_{\ell+1}$ in G and $K \in \{B, R\}$. Then, the* generalized convexity inequality

$$\sum_{t=1}^{\ell+1} z_{K,u_t} - \sum_{t=1}^{\ell} z_{K,v_t} \leq 1 \tag{17}$$

is valid for P_2. In particular, it dominates (11)*, if $\{u_1, \ldots, u_{\ell+1}\} \subseteq V_K$ and $\{v_1, \ldots, v_\ell\} \subseteq V_{\bar{K}}$. Moreover, it is facet-defining, if G is $\mathcal{U}$-arrangeable, where $\mathcal{U} = \{U \subseteq S : |U \cap \{u_1, \ldots, u_{\ell+1}\}| = |U \cap \{v_1, \ldots, v_\ell\}| + 1, H[U] \cap S = U\}$.*

Proof. The proof that (17) is valid for P_2 is similar to the proof of Proposition 15 (see also [10]). If $\{u_1, \ldots, u_{\ell+1}\} \subseteq V_K$ and $\{v_1, \ldots, v_\ell\} \subseteq V_{\bar{K}}$, we can use $y_{v_t} \geq z_{K,v_t}$ and $y_{u_t} \geq 1 - z_{K,u_t}$, given by (12)-(13), to get that (17) dominates (11).

To show facetness, let $K \in \{B, R\}$, $F = \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 \mid \sum_{\ell=1}^{t+1} z_{K,u_\ell} - \sum_{\ell=1}^{t} z_{K,v_\ell} = 1\}$, and $F' = \{(\boldsymbol{y}, \boldsymbol{z}) \in P_2 \mid \boldsymbol{\pi}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} = \pi_0\}$, for some $(\boldsymbol{\pi}, \boldsymbol{\mu}, \pi_0) \in \mathbb{R}^n \times \mathbb{R}^{2n_z} \times \mathbb{R}$. Once more, we show that, if $F \subseteq F'$, then the corresponding underlying hyperplanes are equivalent. First, observe that $(\mathbf{1}, \sum_{i \in H[U]} \boldsymbol{e}^{K,i}) \in F$ for all $U \in \mathcal{U}$, and $\bigcup_{U \in \mathcal{U}} U = S$ due to $\{u_\ell\} \in \mathcal{U}$ and $\{u_\ell, v_\ell, u_{\ell+1}\} \in \mathcal{U}$. Then, since G is $\mathcal{U}$-arrangeable, there is an ordering $i_1, \ldots, i_p$ of $V' = V \setminus S$, $p = |V'|$, such that, for all $k \in \{1, \ldots, p\}$, there exists $U_k \in \mathcal{U}$ satisfying $H[U_k \cup \{i_k\}] \subseteq U_k \cup \{i_1, \ldots, i_k\}$. Then, we proceed similarly to the proof of Proposition 23, by using the points presented in Table 1.

Case	**Points in** F
$\mu_{K,i'} = 0$ for all $i' \in V'$	induction on k with $\left(\mathbf{1}, \sum_{i' \in H[U_k]} \boldsymbol{e}^{K,i'}\right)$, $\left(\mathbf{1}, \sum_{i' \in H[U_k \cup \{i_k\}]} \boldsymbol{e}^{K,i'}\right)$
$\mu_{\bar{K},i'} = 0$ for all $i' \in V$	$\left(\mathbf{1}, \boldsymbol{e}^{K,i}\right)$, $\left(\mathbf{1}, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},i'}\right)$
$\pi_{i'} = 0$ for all $i' \in V_{BR}$	$\left(\mathbf{1}, \boldsymbol{e}^{K,i}\right), \left(\bar{\boldsymbol{e}}^{i'}, \boldsymbol{e}^{K,i} + \boldsymbol{e}^{\bar{K},i'}\right)$, if $i' \in V_{\bar{K}}$ $\left(\mathbf{1}, \boldsymbol{\xi} = \sum_{i \in H[S \cup \{i'\}]} \boldsymbol{e}^{K,i}\right), \left(\bar{\boldsymbol{e}}^{i'}, \boldsymbol{\xi}\right)$, if $i' \in V_K$
$\mu_{K,u_\ell} = \mu_{K,u_{\ell'}}$, $\ell, \ell' \in \{1, \ldots, t+1\}$ $\ell \neq \ell'$	$\left(\mathbf{1}, \boldsymbol{e}^{K,u_\ell}\right), \left(\mathbf{1}, \boldsymbol{e}^{K,u_{\ell'}}\right)$
$\mu_{K,u_\ell} = -\mu_{K,v_\ell}$, $\ell \in \{1, \ldots, t\}$	$\left(\mathbf{1}, \sum_{i \in H[u_\ell, v_\ell, u_{\ell+1}]} \boldsymbol{e}^{K,i}\right), \left(\mathbf{1}, \boldsymbol{e}^{K,u_{\ell+1}}\right)$

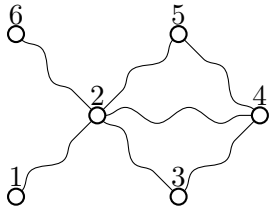
Table 1: Cases in the proof of Proposition 24.

□

It is worth observing that Constraints (15) are special cases of (17) for $\ell = 1$ as well as special cases of (16) for $|S| = 2$ and $j \in D(S)$.

A generalization of inequalities (17) can be obtained by allowing some vertices to appear more than once in the base sequence. In this sense, a *walk* in G is a sequence $\langle v_1, v_2, \ldots, v_\ell \rangle$ of vertices, $\ell \geq 2$, such that $v_i v_{i+1} \in E$ for $i = 1, \ldots, \ell - 1$. The inequalities considered in the next result is illustrated with the example depicted in Figure 5.

Figure 5 – Sequence $\langle u_1 = 1, v_1 = 2, u_2 = 3, v_2 = 4, u_3 = 5, v_3 = 2, u_4 = 6 \rangle$,
$v_1 \in D(u_1, u_2) \cap D(u_1, u_3) \cap D(u_1, u_4) \cap D(u_2, u_3) \cap D(u_2, u_4) \cap D(u_3, u_4)$ and $v_2 \in D(u_2, u_3)$.
Generalized walk inequalities: $z_{K,1} + z_{K,3} + z_{K,5} + z_{K,6} - 2z_{K,2} - z_{K,4} \leq 1$, for $K \in \{B, R\}$.
Snaked segments represent shortest paths of their endpoints.



Proposition 25. *Let* $\langle u_1, v_1, \ldots, u_t, v_t, u_{\ell+1} \rangle$, $\ell \geq 1$, *be a sequence contained in a walk between*

u_1 and $u_{\ell+1}$ in G such that

$$\{v_t, v_{t+1}, \ldots, v_{t'}\} \cap D(u_t, u_{t'+1}) \neq \emptyset, \qquad \forall 1 \leq t \leq t' \leq \ell, \tag{18}$$

and $K \in \{B, R\}$. Then, the generalized walk inequality

$$\sum_{t=1}^{\ell+1} z_{K,u_t} - \sum_{t=1}^{\ell} z_{K,v_t} \leq 1 \tag{19}$$

is valid for P_2.

Proof. Suppose by contradiction that there exists an integer point $(\boldsymbol{y}, \boldsymbol{z}) \in P_2$ that violates (19). Then, there are t, t', $1 \leq t \leq t' \leq \ell$, such that $z_{K,u_t} = z_{K,u'_t} = 1$ and $z_{K,v_{t''}} = 0$ for all $t'' \in [t, t']$, which contradicts (18) and (15). □

Condition (18) requires that, for any two vertices $u_t, u_{t'+1}$ in odd positions in the sequence, there is at least one vertex $v_{t''}$ in an even position between them such that $v_{t''} \in D(u_t, u_{t'+1})$. As a consequence, there cannot be multiple occurrences of a single vertex w in odd positions of the sequence because $D(w, w) = \emptyset$. However, a single vertex can appear multiple times in even positions.

As a counterpart of the generalized C_4 inequalities (8) given for P_1, we now present valid inequalities for P_2 that include variables z for vertices in V_N.

Proposition 26. *Let $S \subseteq V_B \cup V_N$ and $T \subseteq V_R \cup V_N$ such that $|S| = |T| = 2$. If $S \subseteq H[T]$ and $T \subseteq H[S]$, then the following inequality is valid for P_2:*

$$\sum_{i \in S \cap V_N} (1 - z_{B,i}) + \sum_{i \in S \cap V_B} y_i + \sum_{j \in T \cap V_N} (1 - z_{R,j}) + \sum_{i \in T \cap V_R} y_i \geq 2. \tag{20}$$

Proof. Suppose by contradiction that there exists an integer point $(\boldsymbol{y}, \boldsymbol{z}) \in P_2$ that violates (20). Then, by symmetry, we can assume that the first two terms in the left-hand side of (20) are null, which implies $z_{B,i} = 1$ for all $i \in S$. Similarly, since one of the other two terms in the left-hand side is also null, there must be $j \in T$ such that $z_{R,j} = 1$ (with $y_j = 0$ if $j \in V_R$). Since $j \in H[S]$, by (15) we have $z_{B,j} = 1$. If $j \in V_N$, then (14) is violated. If $j \in V_R$, then (12) is violated. Both cases contradict $(\boldsymbol{y}, \boldsymbol{z}) \in P_2$. □

It is worth noting that, unlike the previous cases, in this case we cannot replace all variables y by variables z. Indeed, the inequality derived from (20) by replacing y_i by $1 - z_{K,i}$ for vertices $i \in V_K \cap (S \cup T)$, $K \in \{B, R\}$, is not valid as some of these vertices may be outliers. For instance, if $S \in V_B$, we can have a feasible solution where one vertex in S is colored blue and the other 3 vertices in $S \cup T$ are colored red. This solution would violate the suggested inequality.

5 Algorithms and Computational Experiments

In this section, we describe branch-and-cut algorithms to solve the 2-GC problem. They are based on the formulations discussed in the previous sections and configured with a specific

objective function. The aim of the conducted computational experiments, which results are reported in the sequel, is twofold. From one side, the efficiency of separation algorithms for facet-defining inequalities to solve the corresponding formulations to optimality is analyzed. From the other side, the accuracy of the geodesic approach as a classification model is discussed in some random and realistic instances.

The computational experiments were performed on a machine with an Intel i7-7500 2.7 GHz 4 cores processor, 8 GB RAM, and 64 bits Linux OS. The implementation was made using the programming language `C++` and the `CPLEX` package version 12.9 (with default parameters) to solve the linear and integer programming models.

5.1 Objective Functions

Given a feasible solution (A_B, A_R) of the 2-GC problem, the vertices in $V_{BR} \setminus (A_B \cup A_R)$ are the outliers while those in $V_N \setminus (H[A_B] \cup H[A_R])$ are called *uncovered* (by (A_B, A_R)). Outliers and uncovered vertices establish a discrepancy between (A_B, A_R) and V with respect to linear separability. In this sense, different criteria may be used to calibrate the optimization function of the formulations as a measure of how much V deviates from the linearly separable sets A_B, A_R.

In practice, there are several "centrality measures" that could ground an objective function. The basic one would be the minimization of the outliers. For the compact formulation, we could also introduce variables z in the objective function in order to penalize uncovered vertices.

The experiments presented here consider the minimization of the number of outliers as objective function in both formulations, ILP1 and ILP2. Other evaluated functions have not led to better results in terms of classification measures such as accuracy.

5.2 Pre-processing

Computing all-pairs shortest paths is a straightforward adaptation of breadth-first search (unweighted cases) or Floyd-Warshall's (weighted cases) algorithms.

In the unweighted case, set $D(h, j)$ comprises all internal vertices in all shortest paths from h to j. To calculate all $D(h, j)$ sets for a graph instance $G = (V, E)$, we applied a breadth-first search algorithm for each source $h \in V$. In such an algorithm, each vertex j has its $D(h, j)$ set updated every time an adjacent vertex i of j (i.e., $(i, j) \in E$) with lower distance to h ($\delta(h, i) < \delta(h, j)$) is reached. When it happens, the set $D(h, j)$ is updated in the following way: $D(h, j) \leftarrow D(h, j) \cup \{i\} \cup D(h, i)$. This update takes $O(|V|)$. So, for a given source $h \in V$, the sets $D(h, j)$, for all $j \in V$, are determined in time $O(n + n * m)$. Therefore, the complexity to calculate $D(h, j)$ for all $h \in V$ and $j \in V$ is $O(n^2 + n^2 * m))$. In weighted case, we simply apply Floyd-Warshall's, which takes $O(n * n^3) = O(n^4)$ to calculate these sets, since the updates $D(h, j) \leftarrow D(h, j) \cup \{i\} \cup D(h, i)$ takes $O(|V|)$.

For $k \geq 1$, let $D^k[S]$ be the result of the iterative application of operator D from S for k iterations, i.e. $D^1[S] = D[S]$ and $D^{k+1}[S] = D[D^k[S]]$. Note that $D^{k+1}[S] = D^k[S]$ if and only if $D^k[S]$ is convex. To calculate the convex hull W of a subset S, we start with $W' = S$ and iteratively update it. At each iteration, we add to W' all vertices in $D(u, v)$, for every pair $u, v \in W'$ (with at least one of them added to W' in the previous iteration). This step is repeated

until W' does not change. At this point, we get $W = W'$. Sets $D(u, v)$ can be determined a priori with the BFS-like or Floyd-Warshall's algorithm mentioned above.

Given a graph instance G, we calculate the convex hulls $H[V_B]$ and $H[V_R]$ and do the following. If a vertex $i \in V_N$ neither belongs to the convex hull of V_B nor to the convex hull of V_R, then i can never be reached by any class in a feasible solution. In this case, we fix $z_{B,i} = 0$ and $z_{R,i} = 0$ in ILP2. On the other hand, if a vertex $i \in V_N$ does not belong to the convex hull of V_R but it does belong to the convex hull of V_B, then we know that i can never be reached by the red class in a feasible solution, so we fix $z_{R,i} = 0$ in ILP2. A similar fixing can be done in the case where i belongs to the convex hull of V_R but does not belong to the convex hull of V_B.

5.3 Separation Algorithms

The valid inequalities derived for ILP1 or ILP2 are not directly included in the corresponding formulation. Instead, they are only added if violated by the current relaxed solution. This is accomplished by separation routines.

5.3.1 Generalized C_4 Inequalities Separator

A separation algorithm for inequalities (8) and (20) can be obtained by just enumerating and storing them in a list to check for violation. This list can be created during the construction of the initial integer linear programming model by enumerating all 2-sized subsets of $(V_B \cup V_N)$ and $(V_R \cup V_N)$ and verifying the subset pairs that satisfy the requirements of the corresponding inequality. For the sake of time efficiency, our implementation seeks for pairs $(\{i, i'\}, \{j, j'\})$ such that $j, j' \in D(i, i')$ and $i, i' \in D(j, j')$, instead of pairs such that $j, j' \in H[i, i']$ and $i, i' \in H[j, j']$. This way, we are possibly not considering all inequalities (8) or (20).

In practice, searching this list, instead of enumerating all pairs each time the separator runs, drastically decreases the algorithm's overall running time, since the list size is, on average, much smaller than the worst-case (around 0.01% of $|V(G)|^4$).

5.3.2 Generalized Convexity Inequalities Separator

We developed a clever separation algorithm for the generalized convexity inequalities (17), including the case $t = 1$ which corresponds to constraints (15). The idea is the following: given $h \in V$ and $K \in \{B, R\}$, find a sequence contained in a shortest path from h in G whose corresponding inequality yields the maximum value in the left-hand side of (17). We search for such a path in a dag (direct acyclic graph) composed of all sequences contained in a shortest paths starting at root $h \in V$ to any other vertex in G. To create such a dag for a given vertex $h \in V$, we apply the steps below (to be illustrated in the graph of Figure 6a):

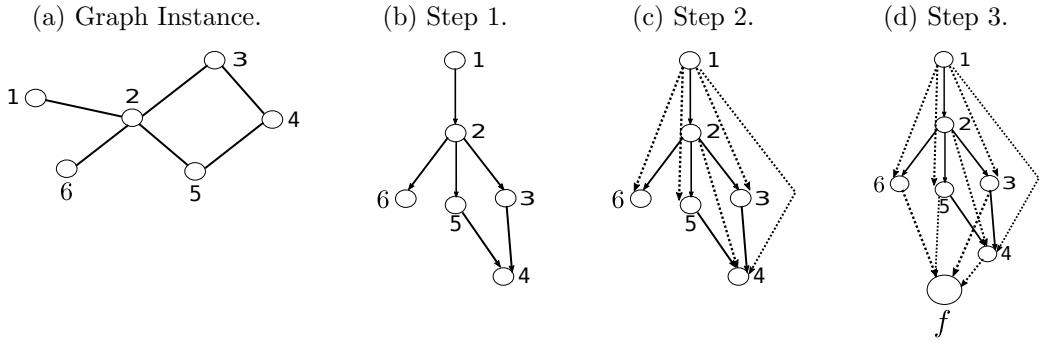
1. Create the dag of all shortest paths starting at h. It is easily calculated by a breadth-first search algorithm similar to the one described in Subsection 5.2 (Figure 6b);

2. Calculate the transitive closure of the dag obtained in Step 1 (a path from h in this transitive dag is a sequence contained in a shortest path starting at h in G). To do so, we apply the known Dijkstra's algorithm (Figure 6c);

3. Finally, create a sink f and add arcs (v, f) for every vertex v of the dag such that $(h, v) \notin E(G)$ (Figure 6d).

The resulting dag, for a given vertex h, is denoted by $Dag[h]$. Observe that the vertices of a path with even cardinality from h to f in $Dag[h]$ gives a sequence, starting at h, that defines a generalized convexity inequality (17).

Figure 6 – Dag used in the separation algorithm (related to vertex 1).



Let us consider a path from h to f with even cardinality in $Dag[h]$ whose sequence of vertices is $S =< h = v_1, v_2, \ldots, v_{2p-1}, v_{2p} = f >$. The left-hand side of the generalized convexity inequality (17) induced by S and class $K \in \{B, R\}$ is

$$z_{K,v_1} - z_{K,v_2} + z_{K,v_3} \ldots - z_{K,v_{2p-2}} + z_{K,v_{2p-1}} =$$
$$\frac{z_{K,v_1}}{2} + \frac{z_{K,v_1} - z_{K,v_2}}{2} - \frac{z_{K,v_2} + z_{K,v_3}}{2} + \ldots - \frac{z_{K,v_{2p-2}} + z_{K,v_{2p-1}}}{2}. \quad (21)$$

Thus, to calculate the most violated path in $Dag[h]$, related to class $K \in \{B, R\}$, we assign to each arc (u, v) of $Dag[h]$ a weight $w(u, v) = \frac{z_{K,u} - z_{K,v}}{2}$. When calculating the weight of a path containing an arc (u, v), we multiply $w(u, v)$ by 1 or -1 depending on whether u appears in an odd or even position in the path, respectively. Therefore, the weight of the path is the sum of the weight of its arcs (with their appropriated signs) plus $\frac{z_{K,h}}{2}$, which yields (21).

We can determine the maximum weighted path by traversing $Dag[h]$ in the topological order and calculating the weight of the paths. We just have to take care to multiply the arc weights by 1 or -1 accordingly. In the end, if the weight of the maximum weighted path is greater than 1, the inequality induced by K and this path is violated; otherwise, no violation exists for K and the sequences starting at h.

The complexity of such a separation algorithm, for a given source h, is mainly determined by the Dijkstra's algorithm, which has complexity $O(|V|^2 \log |V|)$, since the topological sorting algorithm can be implemented within complexity $O(|V| + |E|)$ with a depth-first search algorithm.

5.3.3 Generalized Alternating Path Inequalities Separator

The generalized alternating path inequalities (11) are the counterparts, for ILP1, of the generalized convexity inequalities. Their separation can be obtained by a procedure similar to that one presented in Subsection 5.3.2. In practice, we observed that separating all such inequalities was

not rewarding. Instead, we restricted ourselves to separate the generalized 3-path inequalities (6) (the special case of (11) when $\ell = 1$) by enumeration.

5.3.4 Lazy Constraints Separator for Elementary $\mathcal{N}$-set Inequalities

Since there can be many constraints in ILP1, we designed a lazy constraints algorithm to separate integer solutions. We just search for subsets $S \subseteq V_B$ and $T \subseteq V_R$ whose corresponding $\mathcal{N}$-set elementary constraint is violated. To do so, it is sufficient to transform the current integer solution y of ILP1 into the corresponding integer solution $(\hat{y}, \hat{z})$ of ILP2 by explicitly calculating the convex hull of the basis of each class: $A_K = \{i \in V_K \mid y_i = 0\}$, $K \in \{B, R\}$. Then, we verify:

- If $\hat{y}_i < \hat{z}_{\bar{K}(i),i}$, for some $i \in V_{BR}$, then constraint $(\sum_{j \in A_{\bar{K}(i)}} y_j) + y_i \geq 1$ is violated.
- If $\hat{z}_{B,i} + \hat{z}_{R,i} > 1$, for some $i \in V_N$, then constraint $\sum_{j \in A_B} y_j + \sum_{j \in A_R} y_j \geq 1$ is violated.

To obtain minimal subsets S and T that keep the inequalities violated, we proceed as follows. For the V_{BR}-disjoint constraints, we start with $S = A_{\bar{K}}$ and $T = \{i\}$. If there is $u \in S$ such that $H[S \backslash \{u\}]$ still reaches i, then we remove u from S. This procedure finishes when S becomes minimal.

For the V_N-disjoint constraints, we start with $S = A_B$ and $T = A_R$. If there is $u \in S$ or $w \in T$ such that $H[S \backslash \{u\}] \cap H[T] \cap V_N \neq \emptyset$ or $H[S] \cap H[T \backslash \{w\}] \cap V_N \neq \emptyset$, then we remove u from S or w from T, respectively. This procedure finishes when $S \cup T$ becomes minimal.

It is worth noting that the second type of constraints only needs to be verified if $H[S] \cap T = H[T] \cap S = \emptyset$; otherwise it is covered by the first type. The whole algorithm has complexity $O(|V(G)|^4)$.

5.4 Geodesic Classification Algorithms

The algorithms that we developed for each formulation are branch-and-bound algorithms [18], which implicitly enumerate all feasible solutions of the problem via a decision tree structure. They also use a cutting plane algorithm to solve the linear relaxation of the root node, which includes some valid inequalities (cuts) found by our separation algorithms, and a lazy constraint approach to find feasible integer solutions.

For the set covering formulation ILP1, the main steps of our solution method are described in Algorithm 1. Similarly, the main steps of the solution method for formulation ILP2 are described in Algorithm 2.

5.5 Results and Analysis

We present the computational results of Algorithms 1 and 2 separately for randomly generated instances and for realistic instances.

5.5.1 Random Instances

The random instances used in our experiments were categorized by number of vertices $v \in \{50, 100, 150, 200, 250\}$, graph density percentage $d \in \{5, 10, 20, 30, 50, 70\}$ and initially classified

Algorithm 1: ILP1 solving algorithm

1 Computation of all $D(h, j)$ sets.
2 Initial cutoff: Since a trivial solution is obtained by setting all vertices of a class as outliers, $\min\{|V_B|, |V_R|\}$ is provided as a cutoff.
3 Initial model configuration: All generalized C_4 inequalities (8) (with $D(h, j)$ requirement instead of $H[\{h, j\}]$) are included in the initial model by exhaustive enumeration of all pairs of 2-sized subsets. None of the constraints (4) are used initially.
4 Partial linear relaxation resolution: At the root node of the branch-and-cut tree, we solve the linear relaxation of the initial model together with the generalized 3-path constraints (6) separated as cuts by enumeration.
5 Exact model resolution: Starting from the model obtained after Step 4, we add the integrality constraints and solve the integer formulation by adding minimal (S, T) constraints (4) as lazy constraints.

Algorithm 2: ILP2 solving algorithm

1 Computation of all $D(h, j)$ sets and inclusion of all constraints (12)-(14) in the initial model.
2 Initial cutoff: Since a trivial solution is obtained by setting all vertices of a class as outliers, $\min\{|V_B|, |V_R|\}$ is provided as a cutoff.
3 Initial model configuration: All generalized C_4 inequalities (8) (with $D(h, j)$ requirement instead of $H[\{h, j\}]$) are included in the initial model by exhaustive enumeration of all pairs of 2-sized subsets. None of the constraints (15) are included initially.
4 Partial linear relaxation resolution: At the root node of the branch-and-cut tree, we solve the linear relaxation of the initial model together with inequalities (17) and (20) (with $D(h, j)$ requirement instead of $H[\{h, j\}]$) separated as cuts.
5 Exact model resolution: Starting from the model obtained after Step 4, we add the integrality constraints and solve the integer formulation by adding (15) as lazy constraints by enumeration.

Table 2: Algorithm 1 and Algorithm 2 running times comparison for random instances with $p = 60$ (in seconds).

Instance	Diam	minDg	maxDg	OPT	$T_{ILP1}(s)$	$T_{ILP2}(s)$
v050-d05-p60	9	1	6	12	**0.38**	0.42
v050-d10-p60	4	1	9	14	1.79	**0.71**
v050-d20-p60	3	4	15	15	1.25	**0.54**
v050-d30-p60	3	8	22	15	0.50	**0.38**
v050-d50-p60	2	17	32	15	**0.03**	0.11
v050-d70-p60	2	26	41	15	**0.02**	0.12
v100-d05-p60	6	1	10	29	4.53	**4.21**
v100-d10-p60	4	3	17	29	**3.17**	4.17
v100-d20-p60	3	10	31	30	0.78	**0.71**
v100-d30-p60	2	18	41	30	**0.05**	0.33
v100-d50-p60	2	37	62	30	**0.17**	1.00
v100-d70-p60	2	57	79	30	**0.20**	1.14
v150-d05-p60	5	1	15	44	26.49	**12.22**
v150-d10-p60	3	5	25	45	**2.21**	3.21
v150-d20-p60	3	17	43	45	**0.11**	0.43
v150-d30-p60	2	30	59	45	**0.27**	2.58
v150-d50-p60	2	59	91	45	**0.36**	2.39
v150-d70-p60	2	88	118	45	**0.47**	2.70
v200-d05-p60	4	3	19	60	**21.80**	32.64
v200-d10-p60	3	9	32	60	**0.69**	3.75
v200-d20-p60	3	24	55	60	**0.42**	1.43
v200-d30-p60	2	43	77	60	**0.38**	3.15
v200-d50-p60	2	81	119	60	**1.35**	8.10
v200-d70-p60	2	122	156	60	**1.79**	5.65
v250-d05-p60	4	4	23	75	**13.68**	129.13
v250-d10-p60	3	13	39	75	**0.63**	6.64
v250-d20-p60	2	32	68	75	**0.42**	2.75
v250-d30-p60	2	54	94	75	**1.26**	9.91
v250-d50-p60	2	102	147	75	**4.05**	14.2
v250-d70-p60	2	153	194	75	**5.84**	16.14
AVERAGE	-	-	-	-	**3.17**	9.03

vertices percentage $p \in \{60, 80\}$. The initially classified vertices were distributed equally between the blue and the red classes. For each combination v, d and p, we generated 10 random instances, adding up to 600 random instances overall.

Tables 2-3 present a running time comparison between Algorithm 1 and Algorithm 2 for the random instances. The time limit was set to 3600 seconds ("-" means this time limit was exceeded). The columns are: instance parameters using the format v$< v >$-d$< d >$-br$< br >$ (Instance); diameter of the graph (Diam); minimum degree in the graph (minDg); maximum degree in the graph (maxDg); optimal solution (the minimum number of outliers) (OPT); average running time of Algorithm 1 in seconds ($T_{ILP1}(s)$); average running time of Algorithm 2 in seconds ($T_{ILP2}(s)$).

We also studied the effect of the valid inequalities used in each algorithm. For the sake of comparison, we tested two other versions of each algorithm, each version obtained by the elimination of Step 3 or Step 4, respectively. The observed results are summarized in Tables 4 and 5. They show the performance of the three tested versions with respect to a standard implementation where both steps 3 and 4 were not applied.

We could notice that inequalities (8) were extremely effective: on average, there were $2.6|V|$ inequalities added in ILP1 ($12|V|$ for ILP2 when (20) were added as cuts), and they reduced by 82% (83% for ILP2) the running time and 30% (20% for ILP2) of the number of lazy constraints added. These were the most effective valid inequalities that we found. Remember that inequalities (8) were proved to be facet-defining for the polytope associated with ILP1 (see

Table 3: Algorithm 1 and Algorithm 2 running times comparison for random instances with $p = 80$ (in seconds).

Instance	Diam	minDg	maxDg	OPT	$T_{ILP1}(s)$	$T_{ILP2}(s)$
v050-d05-p80	10	1	6	15	0.47	**0.29**
v050-d10-p80	5	1	10	19	2.24	**1.05**
v050-d20-p80	3	4	16	20	1.61	**0.61**
v050-d30-p80	3	7	21	20	0.24	**0.17**
v050-d50-p80	2	16	31	20	**0.02**	0.09
v050-d70-p80	2	27	41	20	**0.02**	0.13
v100-d05-p80	6	1	10	39	6.42	**4.92**
v100-d10-p80	4	3	18	40	**0.44**	1.09
v100-d20-p80	3	10	29	40	**0.08**	0.39
v100-d30-p80	2	19	42	40	**0.12**	0.70
v100-d50-p80	2	37	62	40	**0.21**	1.41
v100-d70-p80	2	57	80	40	**0.29**	0.82
v150-d05-p80	5	1	14	59	**4.01**	5.39
v150-d10-p80	3	6	25	60	**0.15**	0.60
v150-d20-p80	3	17	43	60	**0.39**	1.12
v150-d30-p80	2	30	59	60	**0.36**	2.18
v150-d50-p80	2	59	90	60	**1.32**	3.31
v150-d70-p80	2	89	119	60	**1.68**	4.41
v200-d05-p80	4	2	19	79	**4.44**	13.65
v200-d10-p80	3	9	33	80	**0.38**	1.28
v200-d20-p80	2	25	56	80	**0.50**	2.93
v200-d30-p80	2	43	77	80	**1.56**	9.76
v200-d50-p80	2	80	118	80	**5.94**	13.43
v200-d70-p80	2	120	156	80	**8.49**	18.47
v250-d05-p80	4	4	22	100	**2.40**	9.48
v250-d10-p80	3	13	38	100	**0.95**	2.73
v250-d20-p80	2	32	67	100	**1.38**	6.51
v250-d30-p80	2	56	95	100	**4.71**	17.12
v250-d50-p80	2	102	146	100	**17.84**	40.88
v250-d70-p80	2	154	193	100	**24.93**	54.93
AVERAGE	-	-	-	-	**3.12**	7.33

Algorithm 1	N. of constraints (6)	N. of inequalities (8)	Lazy Const. Reduction	Time Reduction
Step 3 only	0	$2.6\|V\|$	30%	82%
Step 4 only	$14\|V\|$	0	92%	73%
Step 3 and Step 4	$2.9\|V\|$	$2.6\|V\|$	88%	85%

Table 4: Effect of the valid inequalities for ILP1.

Algorithm 2	N. of inequalities (17)	N. of inequalities (8), (20)	Lazy Const. Reduction	Time Reduction
Step 3 only	0	$12\|V\|$	20%	83%
Step 4 only, with (17)	$8\|V\|$	0	85%	11%
Step 3 and Step 4	$5\|V\|$	$17\|V\|$	79%	87%

Table 5: Effect of the valid inequalities for ILP2.

Section 3.5). However, the generalized C_4 constraints (20) included as cuts in the root node of the branch-and-cut tree showed only a bit improvement of the linear relaxation lower bound in Algorithm 2. It is important to note that we did not obtain good results when all the generalized C_4 constraints (20) were included in the initial model of ILP2.

Constraints (6), added when solving the root node in Step 4 of Algorithm 1, were very effective as well. Adding them as cuts was much better than including all of them in the initial model. On average, there were $14|V|$ of these constraints added in ILP1, and they reduced by 73% the running time and by 92% the number of lazy constraints added.

The generalized convexity inequalities (17), added when solving the root node in Step 4 of Algorithm 2, were not effective in reducing the running time for random instances, although they have reduced by 85% the number of lazy constraints (15) added. On average, there were $8|V|$ generalized convexity inequalities added in ILP2, and they reduced only by 11% the running time, mainly because of its effectiveness on realistic and small random instances (the running time reduction was 34%, and they showed to be very useful for such instances).

The combination of all cited inequalities (i.e., including Steps 3 and 4 in both algorithms) showed an overall running time reduction of 85% for ILP1 and 87% for ILP2, and it drastically reduced the number of lazy constraints added (88% for ILP1 and 79% for ILP2). Moreover, the addition of the generalized C_4 inequalities yielded a very good reduction of inequalities (6) and (17) added in the root node. Despite their theoretical effects, Star tree inequalities (9), (16), and generalized walk inequalities (19) did not reduced the running time, so we did not use them in the final version of the branch-and-cut algorithms.

Regarding the lazy constraints scheme presented in Section 5.3.4, its application in Step 5 was fundamental to reduce the running time (with respect to an implementation with all minimal (S, T) constraints (4) added to the initial model). Actually, it was impractical to solve the problem without the lazy constraints scheme since the number of such constraints is potentially exponential. The same was observed for the lazy constraints of ILP2.

Considering all random instances, the average running time of Algorithm 1 was about a few seconds, which shows to be very good even for medium-sized instances. For a large part of these instances, Algorithm 1 had beaten Algorithm 2 in running time, yielding an overall running time $T_{ILP1}(s)$ smaller than $T_{ILP2}(s)$. The few cases in which Algorithm 2 produced better results lied on instances with a low number of vertices and low density. Overall, $T_{ILP1}(s)$ was better than $T_{ILP2}(s)$ in 49 instance configurations out of 60, which corresponds to 82% of all instances.

Figures 7 and 8 show the running times of Algorithm 1 and Algorithm 2 as a function of the graph density. There are graphics for $p = 60\%$ and $p = 80\%$, where the value of v varies in $\{50, 100, 150, 200, 250\}$. These results show evidences that Algorithm 1 works well, especially for dense medium-sized instances, because in such cases the size of $S \cup T$ for the model constraints is generally smaller, which can reduce the number of constraints. Overall, the instances with $p = 80\%$ or density between 5% and 20% showed to be the hardest to solve.

5.5.2 Realistic Instances

To test the developed algorithms for realistic applications, we performed experiments using instances derived from two realistic datasets, namely Parkinson's disease ([19]) and cardiac Single

Figure 7 – Running time versus density, $p = 60$.

(a) Algorithm 1. (b) Algorithm 2.

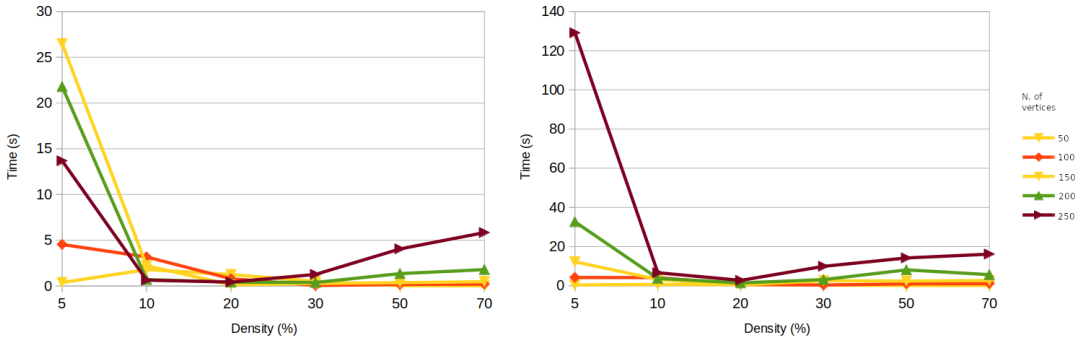
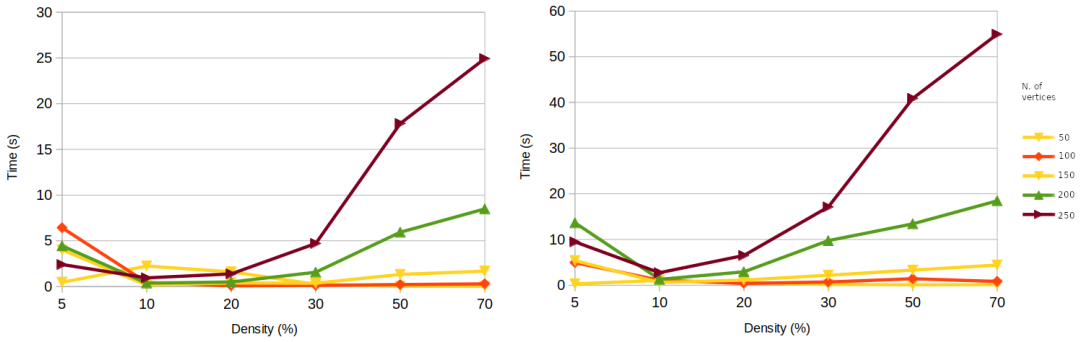


Figure 8 – Running time versus density, $p = 80$.

(a) Algorithm 1. (b) Algorithm 2.



Proton Emission Computed Tomography (SPECT) images ([17]), both available at `https://archive.ics.uci.edu/ml/index.php`. These datasets are used to generate instances of the Euclidean version of the classification problem, where each point represents the information of a patient to be used to predict new diagnostics.

From each dataset, we derived two groups of 10 instances each for the Euclidean classification problem. An instance in the first (resp. second) group is obtained by randomly choosing 20% (resp. 30%) of the points to form the validation set (points used to check the accuracy of a classification algorithm). Then, each instance is transformed into a classification graph (instance of 2-GC) by using the transformation suggested by [25], where each point becomes a vertex. In particular, the validation points correspond to the initially unclassified vertices. As a way to evaluate the class prediction accuracy of our algorithms, we run the well-known *SVM* and *MLP* Euclidean classification algorithms on the Euclidean instances as well as Algorithm 1, Algorithm 2, and the algorithm in [1] on the corresponding 2-GC instances.

The properties of each set of instances is shown in Table 6. The columns are: number of vertices in the classification graph (n); number of edges in the classification graph (m); density of the classification graph (*dens*); diameter of the classification graph (*diam*); minimum degree in the classification graph (*minDg*); maximum degree in the classification graph (*maxDg*).

Tables 7 and 8 show the results for Parkinson and SPECTF instances with $p = 70\%$ ($p = 80\%$ is omitted). The columns are: running time of the algorithm in [1] ($T_{ILPa}(s)$); running time of Algorithm 1 ($T_{ILP1}(s)$); running time of Algorithm 2 ($T_{ILP2}(s)$); accuracy of

Table 6: Properties of each set of instances.

Instance	n	m	density	diam	minDg	maxDg
parkinsons	195	1097	5	10	1	18
spectf	267	1826	5	8	1	36

Table 7: Algorithm in [1], Algorithm 1, Algorithm 2, *SVM*, and *MLP* comparison for Parkinson's instances with $p = 70\%$.

Instance	$T_{ILPa}(s)$	$T_{ILP2}(s)$	$T_{ILP1}(s)$	$Acu_{GC}(\%)$	$Acu_{SVM}(\%)$	$Acu_{MLP}(\%)$
parkinsons-p70-1	695.23	**23.44**	1152.90	**86.21**	63.79	**86.21**
parkinsons-p70-2	14.72	**1.98**	244.76	68.97	68.97	**74.14**
parkinsons-p70-3	123.60	**18.65**	-	77.59	77.59	**81.03**
parkinsons-p70-4	50.88	**1.78**	53.83	**70.69**	**70.69**	**70.69**
parkinsons-p70-5	31.96	**12.64**	35.28	**74.14**	25.86	72.41
parkinsons-p70-6	324.11	**88.87**	-	**75.86**	**75.86**	72.41
parkinsons-p70-7	34.30	**20.96**	392.59	**79.31**	77.59	77.59
parkinsons-p70-8	83.09	**19.83**	-	**75.86**	24.14	**75.86**
parkinsons-p70-9	373.53	**28.23**	-	82.76	**84.48**	81.03
parkinsons-p70-10	67.56	**17.32**	320.20	75.86	**82.76**	75.86
AVERAGE	179.89	**23.37**	-	**76.72**	65.17	**76.72**

the geodesic method ($Acu_{GC}(\%)$) (the best among our three methods); accuracy of the *SVM* method ($Acu_{SVM}(\%)$); accuracy of the *MLP* method ($Acu_{MLP}(\%)$).

Regarding the 20 Parkinson's disease instances, our approach obtained the best accuracy in 10 of them, while 9 and 11 were the scores for *SVM* and *MLP*, respectively. For the 20 SPECT instances, the 2-GC approach presented the best accuracy in 18 instances, while *SVM* and *MLP* did it in 2 and 14 instances, respectively. On average, 2-GC also got the best accuracy, slightly better than the one by *MLP*. Overall, the results show that the accuracy of the 2-GC approach was the best for 28 instances, while *SVM* was the best for only 11 and *MLP* for 25, out of 40 instances.

Comparing the running time of the three algorithms for the geodesic classification in the realistic instances, we notice that Algorithm 2 greatly surpasses the other two algorithms for the Parkinson's instances. On the other hand, for the SPECTF instances, the algorithm in [1] and Algorithm 1, which have nearly equivalent average results, outperform Algorithm 2.

Table 8: Algorithm in [1], Algorithm 1, Algorithm 2, *SVM*, and *MLP* comparison for *SPECT* instances with $p = 70\%$.

Instance	$T_{ILPa}(s)$	$T_{ILP2}(s)$	$T_{ILP1}(s)$	$Acu_{GC}(\%)$	$Acu_{SVM}(\%)$	$Acu_{MLP}(\%)$
spectf-p70-1	**0.17**	2.05	0.25	**83.75**	72.50	**83.75**
spectf-p70-2	0.32	0.79	**0.26**	**78.75**	73.75	**78.75**
spectf-p70-3	**0.15**	1.50	0.16	**80.00**	75.00	**80.00**
spectf-p70-4	0.15	1.65	**0.13**	**76.25**	67.50	75.00
spectf-p70-5	0.18	**0.06**	0.20	**81.25**	41.25	**81.25**
spectf-p70-6	0.08	**0.06**	0.13	**80.00**	62.50	**80.00**
spectf-p70-7	0.20	1.78	**0.19**	**78.75**	51.25	**78.75**
spectf-p70-8	**0.21**	2.60	0.23	**83.75**	60.00	**83.75**
spectf-p70-9	**0.59**	0.75	0.79	**88.75**	87.50	86.25
spectf-p70-10	0.12	**0.06**	0.14	**78.75**	76.25	72.50
AVERAGE	**0.22**	1.13	0.25	**81.00**	66.75	80.00

6 Concluding Remarks

In this work, we studied the 2-class geodesic classification problem (introduced in [1]), a classification problem on graphs that is an analogy of the Euclidean classification problem. This problem presents pure combinatorial optimization aspects and appears as an intersection of a graph convexity problem and the well-known set covering problem. Its applications arise in the fields of data mining and statistics.

We proposed two new integer programming formulations. As the main focus of this work, we studied the polyhedra associated with these formulations, giving some valid inequalities and facet-defining conditions. In order to run computational experiments to validate the accuracy of the geodesic classification approach and the efficiency of the proposed valid inequalities, we developed a branch and cut algorithm for each integer formulation.

The results of the computational experiments show that the proposed solution methods are very promising since the branch-and-cut algorithms proved to be very efficient (in running time and accuracy), even for medium-sized instances. For the random instances, the branch-and-cut based on the set covering formulation was the best one for most of the cases. For the realistic instances, however, the algorithm using the compact formulation seems to present the best computational performance. It is important to remark that the proposed lazy constraints scheme and cutting plane procedure were fundamental to reduce the running time of both algorithms.

We validated the accuracy of the geodesic convexity approach by comparing the prediction provided by the proposed algorithms with two of the most used approaches for the Euclidean convexity classification problem, namely *SVM* and *MLP*. The prediction accuracy of the geodesic approach showed to be stable and as good as such classic linear separation algorithms for the multidimensional space.

We can glimpse some possible directions to enhance and extend this work. As in the Euclidean case, we can use piecewise linear separation instead of linear separation where, for each class, we allow multiple groups that must be pairwise linear separable. This extension has been considered in [14], and the formulations proposed here can be easily adapted to this more general case. Another possible variant is to use edge-weighted classification graphs and calculate convex sets based on minimum weighted paths. The weight of an edge could represent, for instance, how often its extreme vertices are assigned to the same class by different solutions obtained with various resolution methods for the given instance of the *GC* problem. Thus, the weighted edges try to better simulate the underlying pattern of the samples, and a new classification approach based on such a graph may be even more accurate.

We would like to thank all reviewers from the ALIO/EURO Conference who contributed to improve the preliminary version of this work.

References

[1] P. H. M. Araújo, M. Campêlo, R. C. Corrêa, and M. Labbé. The geodesic classification problem on graphs. *Electronic Notes in Theoretical Computer Science*, 346:65 – 76, 2019.

The proceedings of Lagos 2019, the tenth Latin and American Algorithms, Graphs and Optimization Symposium (LAGOS 2019).

[2] B. Aronov, D. Garijo, Y. Núñez-Rodríguez, D. Rappaport, C. Seara, and J. Urrutia. Minimizing the error of linear separators on linearly inseparable data. *Discrete Applied Mathematics*, 160(10):1441 – 1452, 2012.

[3] D. Artigas, S. Dantas, M. C. Dourado, and J. L. Szwarcfiter. Partitioning a graph into convex sets. *Discrete Mathematics*, 311(17):1968–1977, 2011.

[4] E. Balas and S. M. Ng. On the set covering polytope: I. all the facets with coefficients in {0, 1, 2}. *Mathematical Programming*, 43(1):57–69, 1989.

[5] D. Bertsimas and R. Shioda. Classification and regression via integer optimization. *Operations Research*, 55(2):252–271, 2007.

[6] M. Blaum, R. C. Corrêa, J. Marenco, I. Koch, and M. Mydlarz. A set covering approach for the 2-class classification problem. Working paper, 2019.

[7] J. A. Bondy and U. S. R. Murty. *Graph Theory*. Springer, 2008.

[8] R. Buzatu and S. Cataranciuc. Convex graph covers. *The Computer Science Journal of Moldova*, 23(3):251–269, 2015.

[9] M. Campêlo, V. A. Campos, and R. C. Corrêa. On the asymmetric representatives formulation for the vertex coloring problem. *Discrete Applied Mathematics*, 156(7):1097 – 1111, 2008.

[10] M. Campêlo, A. Freire, K. Lima, P. Moura, and Y. Wakabayashi. The convex recoloring problem: polyhedra, facets and computational experiments. *Mathematical Programming*, 156, 01 2015.

[11] R. C. Corrêa, M. Blaum, J. Marenco, I. Koch, and M. Mydlarz. An integer programming approach for the 2-class single-group classification problem. *Electronic Notes in Theoretical Computer Science*, 346:321 – 331, 2019. The proceedings of Lagos 2019, the tenth Latin and American Algorithms, Graphs and Optimization Symposium (LAGOS 2019).

[12] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[13] P. J. Crossno, A. T. Wilson, T. M. Shead, I. Warren L. Davis, and D. M. Dunlavy. Topicview: Visual analysis of topic models and their impact on document clustering. *International Journal on Artificial Intelligence Tools*, 22(05):1360008–1 – 1360008–36, 2013.

[14] P. H. M. de Araújo. *The Geodesic Classification Problem on Graphs*. PhD thesis, PhD thesis, Mestrado e Doutorado em Ciência da Computação, Departamento de Ciência da Computação da Universidade Federal do Ceará, 2019.

[15] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[16] R. M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.

[17] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. S. Goodenday. Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial intelligence in medicine*, 23(2):149–169, Oct 2001.

[18] E. L. Lawler and D. E. Wood. Branch-and-bound methods: A survey. *Oper. Res.*, 14(4):699–719, Aug. 1966.

[19] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, Jun 2007.

[20] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley Interscience Series in Discrete Mathematics and Optimization. Wiley, 1988.

[21] P. M. Pardalos and P. Hansen. *Data mining and mathematical programming*, volume 45. American Mathematical Society, Providence, RI, jan 2008.

[22] I. M. Pelayo. *Geodesic Convexity in Graphs*. Springer-Verlag, New York, 2013.

[23] M. Sánchez-García, M. I. Sobrón, and B. Vitoriano. On the set covering polytope:facets with coefficients in {0, 1, 2, 3}. *Annals of Operations Research*, 81:343–356, Jun 1998.

[24] G. Xu and L. G. Papageorgiou. A mixed integer optimisation model for data classification. *Comput. Ind. Eng.*, 56(4):1205–1215, May 2009.

[25] M. J. Zaki and W. M. Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA, 2014.