Exploring Extensions of Maximum Clique Problems from 2-Graphs to k-Uniform Hypergraphs

Dorian Gailhard

Abstract

My end-of-study internship took place from the 1st of May to the 20th of October at the University of Padua, specifically within the Department of Applied Mathematics. I had the opportunity of working under the guidance of Mr. Francesco Rinaldi, with additional support from one of his former Ph.D. students, Mr. Damiano Zeffiro. The primary objective of this internship was to generalize techniques for solving clustering problems in hypergraphs, with a particular focus on the Maximum Clique Problem (MCP) and various related relaxations.

The report is divided into three main parts, each corresponding to a significant project within my internship. These projects encompass work I have completed and ongoing research. You can access the LaTeX files for each part via the following links: Part 1, Part 2, and Part 3. Furthermore, I've provided a Python implementation of the first part of my internship, which focuses on the Maximum s-defective Clique Problem in hypergraphs. You can find it here.

Contents

1 Introduction

In recent years, the rapid growth of data availability has posed new challenges and opportunities in various fields, from image analysis and bioinformatics to marketing and social network analysis. The ability to extract meaningful insights and patterns from large and complex datasets has become a crucial endeavor. Data clustering, a fundamental technique in unsupervised learning, plays a pivotal role in uncovering hidden structures within data by grouping similar data points together.

The clique model, dating back at least to [24] about social networks, refers to subsets where every elements is in a direct relation to all of the others. The problem of finding maximal cliques has a huge field of applications in domains including telecommunication networks, biochemistry, financial networks, and scheduling (see [38] and [26]), and is NP-hard [21]. However, owing in part to its wide applicability, a large variety of both heuristic and exact approaches has been investigated (see Bomze et al. [4] for a thorough overview of formulations and algorithms going up to 1999; a more recent survey of algorithms is given in Wu and Hao [37]). The problem has been extended to hypergraphs, which can model relations between more than two elements, and numerous heuristics has been proposed and tested on real world data for the finding of maximum cliques in this context (see e.g. [11], [33], [36], [40]).

A significant number of the solution methods proposed (for example, Bomze [1], Bomze et al. [3], Gibbons et al. [14], Kuznetsova and Strekalovsky [22], Motzkin and Straus [25], Pelillo [29], and Pelillo and Jagota [30]) are based on solving the following well-known continuous quadratic programming formulation of the MCP due to Motzkin and Straus [25], where characteristic vectors of maximal cliques are part of the set of local maximizers :

$$\max_{\text{s.t. } x\in\Delta} x^T Ax \tag{1}$$

where Δ is the n-dimensional simplex and $A = (a_{ij})_{i,j\in V}$ denotes the adjacency matrix of the considered graph.

However, one known drawback of this approach in practice is the presence of "infeasible" or "spurious" local maximizers of (1), which are not characteristic vectors for cliques and from which a clique cannot be recovered through any simple transformation. Such points are an undesirable property of the program because they can cause continuous-based heuristics to fail by terminating without producing a clique. In Bomze [1], the author addresses this issue by introducing the following regularized formulation (with $\alpha = 1/2$):

$$\max_{\text{s.t. } x\in\Delta} x^T Ax + \alpha\|x\|^2 \tag{2}$$

In contrast to (1), the local maximizers of (2) have been shown to be in one-one correspondence with the maximal cliques in G [1], and a replicator dynamics approach to solving (2) was shown to reduce the total number of algorithm failures by 30% compared with a similar approach to solving (1) [3]. In Bomze et al. [3], the authors enhanced the algorithm of Bomze [1], adding an annealing heuristic to obtain even stronger results. In addition, it was shown that the correspondence between the local/global optima of (2) and the MCP is maintained for any $\alpha \in (0, 1)$. A similar formulation and approach [2] has also been applied successfully to a weighted version of the MCP. A generalization to hypergraphs was introduced in Rota-Bulò and Pelillo [32].

This continuous formulation served as the central point of my internship, spanning three axes. First, we generalized to hypergraphs the results concerning a continuous formulation of a relaxed clique model that allows up to s missing edges. Then, we extended the continuous formulation to a broader class of regularizers (originally $\frac{1}{2}\|\cdot\|_2^2$ for $k = 2$). Lastly, we studied techniques for converting discrete problems into continuous ones and attempted to apply them to the MCP in hypergraphs.

2 Common definitions and notations

A k-graph is a pair $G = (V, E)$, where $V = [1, \ldots, n]$ is a finite set of vertices and $E \subseteq \binom{V}{k}$ is a set of k-subsets of V, each of which is called a hyperedge. 2-graphs are typically called graphs. We denote by $\overline{G} = (V, \overline{E} = \binom{V}{k} \setminus E)$ the complementary of graph G.

The lagrangian of graph G L_G is defined such that $L_G(x) = \sum_{e \in E} \prod_{i \in e} x_i$.

We denote $\Delta = \{x \in \mathbb{R}^n | x \geq 0 \text{ and } \mathbf{1}^T x = 1\}$ the n-dimensional simplex.

We denote $\mathbf{1}_C$ the characteristic function with logical expression C i.e. $\mathbf{1}_C = 0$ when C is false and $\mathbf{1}_C = 1$ when C is true, and $\mathbf{e}_i$ the i-th column of the identity matrix i.e. a vector with a 1 at index i and 0 everywhere else.

A characteristic vector of a subset A is defined by $x^{(A)} = \left(\frac{\mathbf{1}_{i \in A}}{|A|}\right)_i \in \Delta$.

For every $x \in \Delta$, $i \in [1, \ldots, n]$, the multiplier function $\lambda_i : \Delta \to \mathbb{R}$ for a function f is defined as

$$\lambda_i(x) := \nabla f(x)^T (\mathbf{e}_i - x) \tag{3}$$

or in vector form

$$\lambda(x) := \nabla f(x) - x^T \nabla f(x) \mathbf{1} \tag{4}$$

They condense the KKT conditions in a single line.

A clique is a fully connected subset of vertices. A clique is said to be maximal if it is not contained in any other clique, while it is called maximum if it has maximum cardinality. Finding the largest clique in a graph is known as the MCP, which is a combinatorial problem.

3 A continuous formulation for the Maximum s-defective Clique Problem in k-uniform hypergraphs

3.1 Introduction

Since the strict requirement that every two elements have a direct relation is often not satisfied in practice (for example due to experimental errors when working on proteins, see [39] where the concept of s-defective clique was introduced), many relaxations of the clique model have been proposed (see, e.g., [28] for a survey). In this part we are interested in s-defective cliques (see also for example [35] and [27]), where up to s links are allowed to be missing.

In [9], following a long history of continuous formulations for this kind of problems, the authors defined a regularized version of a cubic continuous formulation for the Maximum s-defective Clique Problem (MsdCP) proposed in [34], and then applied variants of the classic Frank–Wolfe (FW) method (see [12]) to this formulation.

The first part of my internship was dedicated to the search of a continuous formulation for the MsdCP in hypergraphs, inspired by existing formulations for the simpler case $k = 2$, and proving its equivalence with the combinatorial problem. Then I generalized additional results about two algorithms able to solve the continuous formulation.

This part is organized under the same structure as [9] : after giving some basic notations in subsection 3.2, previous works on the topic are summarized before studying the regularized maximum s-defective clique formulation for hypergraphs in subsection 3.3, which is easily extended to s-plexes in subsection 3.4. Then, in subsection 3.5, the results obtained for the FDFW algorithm and the FW variant tailored to the MsdCP that can be found in [9] are generalized. Finally in subsection 3.7, preliminary numerical comparisons between the FDFW and the FWdc on our formulation are reported, with the implementation used available here.

3.2 Additional definitions and notations specific to this section

We denote by $G(y) := (V, E(y))$ the weighted graph which has the vertices of $\text{supp}(y)$ with weights equal to y_e, where y_e is the coefficient in y coding for edge e.

We extend the definition of the lagrangian to weighted graphs of the form $G(y)$ by defining $L_{G(y)}$ such that $L_{G(y)}(x) := \sum_{e \in E} y_e \prod_{i \in e} x_i$.

The search spaces in the following problems will be defined later and will always be denoted $\mathcal{P}_s$.

For $p \in \mathcal{P}_s$ we define as $T_{\mathcal{P}_s}(p) := \{v - p | v \in \mathcal{P}_s\}$ the cone of feasible directions at p in $\mathcal{P}_s$, while for $r \in \mathbb{R}^{|V|+|\overline{E}|}$ we define $T^0_{\mathcal{P}_s}(p, r)$ as the intersection between $T_{\mathcal{P}_s}(p)$ and the plane orthogonal to r :

$$T^0_{\mathcal{P}_s}(p, r) := \left\{ d \in T_{\mathcal{P}_s}(p) | d^T r = 0 \right\}$$

3.3 A continuous characterization of maximal s-defective cliques in k-graphs

3.3.1 Specific definitions and notations

A clique is a fully connected subset of vertices. A clique is said to be maximal if it is not contained in any other clique, while it is called maximum if it has maximum cardinality. An s-defective clique is a relaxation of this definition allowing up to s missing edges in the clique. Finding the largest s-defective clique in a graph is known as the MsdCP, which is a combinatorial problem.

Let $\mathcal{D}_s(G) := \{y \in [0, 1]^{\overline{E}} | \mathbf{1}^T y \leq s\}$ and $\mathcal{P}_s := \Delta \times \mathcal{D}_s(G)$. The rest of the subsection is dedicated to the derivation of a continuous formulation of the MsdCP.

3.3.2 Previous work

The first step of my internship was to get familiar with [34], [9] and [10] which will be summarized in this section.

Turàn's theorem states that :

$$\max_{\text{s.t. } x \in \Delta} x^T A x = 1 - \frac{1}{\omega(G)} \tag{5}$$

where $\Delta = \{x \in \mathbb{R}^n | x \geq 0 \text{ and } 1^T x = 1\}$, A is the adjacency matrix of the graph, and $\omega(G)$ is the size of the maximum clique of the graph. Characteristic vectors of maximum cliques are global maxima of the objective function, but some global maxima are not linked to maximum cliques (these maximizers are called "spurious" solutions). To address it, in [5] a regularization was introduced to suppress those spurious maxima :

$$\max_{\text{s.t. } x \in \Delta} x^T A x + \alpha \|x\|_2^2 \tag{6}$$

with $\alpha \in (0, 1)$. The global and local maximizers of this problem are then the characteristic vectors of respectively maximum and maximal cliques.

The continuous unregularized formulation for the MCP was generalized in [34] to the MsdCP :

$$\max_{\text{s.t. } (x,y) \in \Delta \times \mathcal{D}_s(G)} x^T (A + A(y)) x \tag{7}$$

where $\Delta = \{x \in \mathbb{R}^n | x \geq 0 \text{ and } 1^T x = 1\}$ and $\mathcal{D}_s(G) = \{y \in \{0, 1\}^{\overline{E}} | 1^T y \leq s\}$ (which is the set of all the possible choices of s additional "fake" edges to complete the graph). Here, $A(y)$ is the adjacency matrix of the graph made of the s selected fake edges. In the end, the problem can be formulated as such : finding the s best edges to add to the original graph to have the largest clique in the augmented graph. $y \in \{0, 1\}^{\overline{E}}$ in the definition of $\mathcal{D}_s(G)$ can be relaxed to $y \in [0, 1]^{\overline{E}}$ in order to obtain a fully continuous formulation.

Some maximizers of the previous problem are spurious, and like for the MCP it is possible to add a regularization to suppress those maxima, which was done in [9] :

$$\max_{\text{s.t. } (x,y)\in\Delta\times\mathcal{D}_s(G)} x^T(A+A(y))x + \frac{\alpha}{2}\|x\|^2 + \frac{\beta}{2}\|y\|^2 \tag{8}$$

where $\alpha \in (0,2)$ and $\beta > 0$. This formulation ensures that the local and global maximizers of the above problems are strictly characteristic vectors of respectively maximal and maximum s-defective cliques of the graph. The authors also introduced a variant of the Frank-Wolfe (FW) algorithm adapted to this formulation and proved various convergence results.

Finally, the regularized maximum clique formulation was generalized to k-uniform hypergraphs in [10] :

$$\min_{\text{s.t. } x\in\Delta} \sum_{e\in\overline{E}} \prod_{i\in e} x_i + \tau\|x\|_k^k \tag{9}$$

where $0 < \tau \le \frac{1}{k(k-1)}$ (with strict inequality when $k = 2$). Beware that contrary to the previous formulations that searched for a maximum on the graph, this one searches for a minimum on the **complement** of the graph (the sum is indexed on $\overline{E}$ and not E).

3.3.3 Formulation found

Combining the ideas in these works, I came to the continuous formulation described in this part.

Let $G := (V, E)$ be a k-graph with vertices V and edges E. Let $0 < \alpha \le \frac{1}{k(k-1)}$ (with strict inequality for $k = 2$) and $\beta > 0$ and consider the following problem :

$$\min_{\text{s.t. } (x,y)\in\mathcal{P}_s} L_{\overline{G}}(x) - L_{G(y)}(x) + \alpha\|x\|_k^k - \beta\|y\|_2^2 \tag{10}$$

In the following, we will denote $h(x,y) := L_{\overline{G}}(x) - L_{G(y)}(x) + \alpha\|x\|_k^k - \beta\|y\|_2^2$. We claim that minimizers of this problem are attained at $p = (x^{(C)}, y^{(p)})$, where $s \ge l = \mathbf{1}^T y^{(p)} \in \mathbb{N}$, with C an l-defective clique in G which is also a maximal clique in $G \cup G(y^{(p)})$, and $y^{(p)} \in \{0,1\}^{\overline{E}}$ such that $y_e^{(p)} = 1$ for every $e \in \binom{C}{k} \cap \overline{E}$ and with $\operatorname{supp}(y^{(p)})$ of maximum cardinality under these constraints, and conversely we also claim that every maximal s-defective clique in G is associated to such a minimizer.

Remark 3.3.1. *To explain briefly the motivation behind this formulation, it comes pretty straightforwardly from the combination of the ideas developed in [10] (namely the objective function in (9), and that it is necessary to search for a minimum on the complement of the graph instead of a maximum on the graph itself) and in [9] (namely that the MsdCP can be seen as two nested problems : first finding the best s fake edges to add to the graph and then reducing the problem to a standard MCP on the augmented graph with an additional regularization which is visible in (8)).*

These ideas suggest the following formulation :

$$\min_{s.t.\ (x,y)\in\mathcal{P}_s} L_{\overline{G\cup G(y)}}(x) + \alpha\|x\|_k^k \pm \beta\|y\|_l^l \tag{11}$$

where $\pm$ and l are to be chosen later.

Then it seems natural to develop $L_{\overline{G\cup G(y)}}(x) = L_{\overline{G}\setminus G(y)}(x) = L_{\overline{G}}(x) - L_{G(y)}(x)$. At this point $L_{G(y)}(x)$ only has a meaning when $y \in \{0,1\}^{\overline{E}}$, and it seems logical to define it as $L_{G(y)}(x) := \sum_{e\in\overline{E}} y_e \prod_{i\in e} x_i$ in order to obtain something behaving similarly to $x^T A(y) x$ in [9]. This is even more obvious if we notice that the lagrangian can be expressed as a tensor product by $L_G(x) = \mathcal{A}x^k$ where $\mathcal{A}$ is the adjacency tensor of G. Then $\pm$ and l can be chosen to make the reasoning in [9] works.

Recall that in our polytope-constrained setting, (second order) sufficient conditions for the local minimality of $p \in \mathcal{P}_s$ are

$$\nabla h(p)^T d \geq 0 \text{ for all } d \in T_{\mathcal{P}_s}(p) \tag{12}$$

and

$$d^T \nabla^2 h(p) d > 0 \text{ for all } d \in T^0_{\mathcal{P}_s}(p, \nabla h(p)) \tag{13}$$

Recall also that we have the following result (given by [10]) :

Theorem *Let G be a k-graph and $0 < \alpha \leq \frac{1}{k(k-1)}$ (with strict inequality for $k = 2$). A vector $x \in \Delta$ is a local (global) minimizer of (10) if and only if it is the characteristic vector of a maximal (maximum) clique of G.*

3.3.4 Useful formulas

$$L_{\overline{G}}(x) - L_{G(y)}(x) = \sum_{e \in \overline{E}} \prod_{i \in e} x_i - \sum_{e \in E(y)} y_e \prod_{i \in e} x_i \tag{14}$$

When $y \in \{0,1\}^{\overline{E}}$, we can rewrite

$$L_{\overline{G}}(x) - L_{G(y)}(x) = L_{\overline{G} \setminus G(y)}(x) = L_{\overline{G \cup G(y)}}(x) \tag{15}$$

The first and second order derivatives of h are the following :

$$\frac{\partial h}{\partial x_j}(x,y) = \sum_{e \in \overline{E}} \mathbf{1}_{j \in e} \prod_{i \in e \setminus \{j\}} x_i - \sum_{e \in E(y)} y_e \mathbf{1}_{j \in e} \prod_{i \in e \setminus \{j\}} x_i + \alpha k x_j^{k-1} \tag{16}$$

$$\frac{\partial^2 h}{\partial x_i \partial x_j}(x,y) = \mathbf{1}_{i \neq j} \left[\sum_{e \in \overline{E}} \mathbf{1}_{i,j \in e} \prod_{l \in e \setminus \{i,j\}} x_i - \sum_{e \in E(y)} y_e \mathbf{1}_{i,j \in e} \prod_{l \in e \setminus \{i,j\}} x_i \right] + \mathbf{1}_{i=j} \alpha k(k-1) x_i^{k-2} \tag{17}$$

$$\frac{\partial h}{\partial y_e}(x,y) = -\prod_{i \in e} x_i - 2\beta y_e \tag{18}$$

$$\frac{\partial^2 h}{\partial y_e \partial y_{e'}}(x,y) = -2\beta \mathbf{1}_{e=e'} \tag{19}$$

For readers that are not familiar with tensors, we advise reading [23], but knowledge of tensors are necessary only for convergence results at the end of the section. Denote $\mathcal{A}(G)$ the adjacency tensor of graph $G = (V, E)$, defined such that

$$\forall e \in \binom{|V|}{k}, \ (\mathcal{A}(G))_e := \mathbf{1}_{e \in E} \tag{20}$$

which is a symmetric tensor. Denoting $e = (i_1, \ldots, i_k)$, we can express the Lagrangian of graph G and its derivatives at $x \in \Delta$ with means of tensor theory :

$$L_G(x) = \frac{1}{k!} \langle \mathcal{A}(G), \underbrace{x \circ \cdots \circ x}_{\text{k times}} \rangle = \frac{1}{k!} \sum_{i_1, \ldots, i_k = 1}^{|V|} \mathcal{A}(G)_e \prod_{j \in e} x_j = \frac{1}{k!} \mathcal{A}(G) x^k \tag{21}$$

$$\nabla L_G(x) = \frac{1}{(k-1)!} \left(\sum_{i_1, \ldots, i_{k-1} = 1}^{|V|} \mathcal{A}(G)_e \prod_{i \in e \setminus \{i_k\}} x_i \right)_{1 \leq i_k \leq |V|} = \frac{1}{(k-1)!} \mathcal{A}(G) x^{k-1} \tag{22}$$

$$\nabla^2 L_G(x) = \frac{1}{(k-2)!} \left(\sum_{i_1, \ldots, i_{k-2} = 1}^{|V|} \mathcal{A}(G)_e \prod_{i \in e \setminus \{i_k, i_{k-1}\}} x_i \right)_{1 \leq i_k, i_{k-1} \leq |V|} = \frac{1}{(k-2)!} \mathcal{A}(G) x^{k-2} \tag{23}$$

where $\circ$ is the outer product.

3.3.5 Equivalence with a previously derived formulation for the case $k = 2$

Here we provide a short proof of the equivalence between the new formulation in the case $k = 2$ and formulation (8) derived in [9].

For $x \in \Delta$,

$$1 = \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j = 2 \sum_{i<j} x_i x_j + \sum_{i=1}^{n} x_i^2 \tag{24}$$

$$= 2 \sum_{[i,j]\in E} x_i x_j + 2 \sum_{[i,j]\in \overline{E}} x_i x_j + \sum_{i=1}^{n} x_i^2 \tag{25}$$

$$= 2L_G(x) + 2L_{\overline{G}}(x) + \|x\|_2^2 \tag{26}$$

From where we obtain

$$L_{\overline{G}}(x) = \frac{1}{2} - L_G(x) - \frac{1}{2}\|x\|_2^2 \tag{27}$$

Thus

$$\min_{(x,y)\in\mathcal{P}_s} \left[L_{\overline{G}}(x) - L_{G(y)}(x) + \alpha\|x\|_2^2 - \beta\|y\|_2^2\right] \tag{28}$$

$$= \min_{(x,y)\in\mathcal{P}_s} \left[\frac{1}{2} - L_G(x) - L_{G(y)}(x) - \frac{1}{2}\|x\|_2^2 + \alpha\|x\|_2^2 - \beta\|y\|_2^2\right] \tag{29}$$

$$= \frac{1}{2} - \max_{(x,y)\in\mathcal{P}_s} \left[L_{G\cup G(y)}(x) + \frac{1}{2}\|x\|_2^2 - \alpha\|x\|_2^2 + \beta\|y\|_2^2\right] \tag{30}$$

So for $\alpha \in (0, \frac{1}{2})$ and $\beta > 0$, we recover the initial formulation.

3.3.6 Correctness of the formulation in general

Thankfully the proof of equivalence between the continuous and discrete formulations for the case $k = 2$ of [9] was easily generalizable to $k > 2$. After some suggestions by Mr. Zeffiro about some reasonings that could be simplified, we arrived to the proof described in this part.

Proposition 3.3.2 (Characterization of local maxima for h)**.** *The following are equivalent :*

(i) $p = (x, y) \in \mathcal{P}_s$ is a local minimizer for $h(x, y)$

(ii) p is a strict local minimizer

(iii) $p = (x^{(C)}, y^{(p)})$, where $s \geq l = \mathbf{1}^T y^{(p)} \in \mathbb{N}$, with C an l-defective clique in G which is also a maximal clique in $G \cup G(y^{(p)})$, and $y^{(p)} \in \{0,1\}^{\overline{E}}$ such that $y_e^{(p)} = 1$ for every $e \in \binom{C}{k} \cap \overline{E}$ and with supp$(y^{(p)})$ of maximum cardinality under these constraints.

In either of these equivalent cases, we have

$$h(p) = \alpha|C|^{1-k} - \beta l \tag{31}$$

Proof. Let $p := (x^{(p)}, y^{(p)}) \in \mathcal{P}_s$.

$(ii) \Rightarrow (i)$: if p is a strict minimum, it is obviously a minimum.

$(i) \Rightarrow (iii)$: If p is a minimizer of h, then in particular $y^{(p)}$ must be a minimizer of $h(x^{(p)}, \cdot)$ which is strictly concave (indeed $\frac{\partial^2 h}{\partial y_e \partial y_{e'}}(x, y) = -2\beta\mathbf{1}_{e=e'}$ so its hessian is definite negative), thus y must be a vertex of $[0,1]^{\overline{E}}$ i.e. $y \in \{0,1\}^{\overline{E}}$.

Then $x^{(p)}$ must also be a local minimizer for $h(\cdot, y^{(p)})$, which is (up to a constant) a regularized maximal clique relaxation for the augmented graph $G \cup G(y^{(p)})$.

The theorem recalled in preamble then gives us that the local minimizers of this function are $x = x^{(C)}$ with C a maximal clique in $G \cup G(y^{(p)})$. Since $G \cup G(y^{(p)})$ is defined by adding s fake edges to G, C is necessarily an s-defective clique in G.

$(iii) \Rightarrow (ii)$: For a fixed $p = (x^{(C)}, y^{(p)})$ with C, $y^{(p)}$ satisfying the conditions of point (iii).

Case 1 : $C = V$ i.e. the clique is the whole graph (possible if the graph misses less than s edges and the budget is enough to fully complete it). Thus necessarily $y^{(p)} = \mathbf{1}$ i.e. the vector with 1 everywhere and $x^{(C)} = x^{(V)}$ i.e. the characteristic vector for the whole graph. Denote $l = \mathbf{1}^T y^{(p)}$. Then

$$h(x^{(C)}, y^{(p)}) = \alpha \left(\frac{1}{|V|}\right)^{k-1} - l\beta \tag{32}$$

Assume there exists a minimizer $z' = (x', y') \neq (x, y)$. By $(i) \Rightarrow (iii)$, z' satisfies the conditions of (iii), and there is at least $a \geq 1$ missing edges in $G \cup G(y')$ and $b \geq 1$ missing nodes in $\mathrm{supp}(x')$ which is a clique. So

$$\begin{aligned} h(x', y') &= \alpha \left(\frac{1}{|V| - b}\right)^{k-1} - (l - a)\beta & (33)\\ &> \alpha \left(\frac{1}{|V|}\right)^{k-1} - l\beta & (34)\\ &= h(x^{(C)}, y^{(p)}) & (35) \end{aligned}$$

This is true for every minimizer, so in particular this is true for every element of $\mathcal{P}_s$, i.e. $p = (x^{(C)}, y^{(p)})$ is a strict minimizer.

Case 2 : $C \neq V$ i.e. at least one node is not in the clique. Let $\overline{C} := V \setminus C$, $S := \mathrm{supp}(y^{(p)})$, $\overline{S} := \overline{E} \setminus S$, $g := \nabla h(p)$ and $H := \nabla^2 h(p)$.

For every $i \in V$ we have

$$\begin{aligned} g_i = \frac{\partial h}{\partial x_i}(x, y) &= \sum_{e \in \overline{E}} \mathbf{1}_{i \in e} \prod_{j \in e \setminus \{i\}} x_j^{(C)} - \sum_{e \in E(y)} \mathbf{1}_{i \in e} y_e \prod_{j \in e \setminus \{i\}} x_j^{(C)} + \alpha k (x_i^{(C)})^{k-1} \\ &= \sum_{e \in \overline{E} \setminus E(y)} \mathbf{1}_{i \in e} \prod_{j \in e \setminus \{i\}} x_j^{(C)} + \alpha k (x_i^{(C)})^{k-1} \end{aligned} \tag{36}$$

In particular, for $i \in C$

$$g_i = \alpha k \left(\frac{1}{|C|}\right)^{k-1} \tag{37}$$

because the product is non-null only if for all x_j in the edge except x_i, $x_j = \frac{1}{|C|}$, i.e. only if all $j \in C$, so because $i \in C$ then the edge would be in the clique so not in $\overline{E} \setminus E(y)$.

For every $i \in \overline{C}$

$$g_i = \sum_{e \in \overline{E} \setminus E(y)} \mathbf{1}_{i \in e} \prod_{j \in e \setminus \{i\}} x_j^{(C)} \geq \left(\frac{1}{|C|}\right)^{k-1} > \alpha k \left(\frac{1}{|C|}\right)^{k-1} \tag{38}$$

because first by hypothesis there is at least one missing edge between $k-1$ nodes of the clique and one missing node (or else the node could be added to the clique which would not be maximal), then $\alpha \leq \frac{1}{k(k-1)}$ with strict inequality for $k = 2$ so $\alpha k < 1$.

For $e \in \overline{E}$ we have

$$g_e = \frac{\partial h}{\partial y_e}(x, y) = -\prod_{i \in e} x_i^{(C)} - \beta k (y_e^{(p)})^{k-1} \tag{39}$$

and in particular $g_e = 0$ for $e \in \overline{S}$ (because then $y_e = 0$ and at least one $x_i = 0$ or else the edge would be in the clique and so in $E \cup S$), while for $e \in S$

$$g_e = -\left(\frac{1}{|C|}\right)^k - \beta k < 0 \tag{40}$$

Let d be a feasible direction from p so that $d := v - p$ with $v \in \mathcal{P}_s$. Let $\sigma_S := \sum_{e \in S} g_e$, $\sigma_C := \sum_{i \in C} v_i = 1 - \sum_{i \in \overline{C}} v_i \in [0, 1]$, and $m_{\overline{C}} := \min_{i \in \overline{C}} g_i$.

$$\begin{aligned}
g^T p &= \sum_{i \in C} x_i^{(C)} g_i + \sum_{e \in \overline{E}} y_e^{(p)} g_e + \sum_{i \in \overline{C}} x_i^{(C)} g_i & (41)\\
&= \frac{1}{|C|} \sum_{i \in C} g_i + \sum_{e \in S} g_e & (42)\\
&= \alpha k \left(\frac{1}{|C|}\right)^{k-1} + \sigma_S & (43)
\end{aligned}$$

Denoting by g_V the part of the gradient dedicated to the vertices, we also have :

$$g_V^T v_V = g_C^T v_C + g_{\overline{C}}^T v_{\overline{C}} \tag{44}$$

First,

$$g_C^T v_C = \alpha k \left(\frac{1}{|C|}\right)^{k-1} \sum_{i \in C} v_i = \alpha k \left(\frac{1}{|C|}\right)^{k-1} \sigma_C = \alpha k \left(\frac{1}{|C|}\right)^{k-1} \left(1 - \sigma_{\overline{C}}\right) \tag{45}$$

Then,

$$g_{\overline{C}}^T v_{\overline{C}} \geq m_{\overline{C}} \sigma_{\overline{C}} \geq \alpha k \left(\frac{1}{|C|}\right)^{k-1} \sigma_{\overline{C}} \tag{46}$$

Thus,

$$g_V^T v_V \geq \alpha k \left(\frac{1}{|C|}\right)^{k-1} \tag{47}$$

We also have

$$g_{\overline{E}}^T v_{\overline{E}} = g_S^T v_S + g_{\overline{S}}^T v_{\overline{S}} = g_S^T v_S \geq \sigma_S \tag{48}$$

because $g_e < 0$ and $v_e \leq 1$ for every $e \in S$.

In the end,

$$g^T d = g_V^T v_V + g_{\overline{E}}^T v_{\overline{E}} - g^T p \geq 0 \tag{49}$$

We have equality iff there is equality in (46) and (48), thus iff $v = (x^{(v)}, y^{(v)})$ with $\operatorname{supp}(x^{(v)}) \subset C$ (as $\alpha k < 1$ it is impossible to have an equality if $\sigma_{\overline{C}} \neq 0$) and $y^{(v)} = y^{(p)}$ (we need $v_e = 1$ for each $e \in S$ and then all the budget has been used if $s \leq \overline{E}$ or there is simply nowhere to spend it in the contrary case). In particular p is a first order stationary point with

$$T^0_{\mathcal{P}_s}(p, g) = \left\{d \in T_{\mathcal{P}_s}(p) | d = v - p, v_{\overline{C}} = 0, v_{\overline{E}} = p_{\overline{E}}\right\} = \left\{d \in T_{\mathcal{P}_s}(p) | d_{\overline{C}} = d_{\overline{E}} = 0\right\} \tag{50}$$

Now denote by H_C the submatrix of the hessian with indices in C. The objective function here is the same (up to a constant) as for [10], which gives us

$$H_C = \alpha k (k-1) \left(\frac{1}{|C|}\right)^{k-2} \mathbf{I} > 0 \tag{51}$$

This proves the claim since we have sufficient conditions for local minimality. $\square$

As a corollary, the global optimum of h is achieved on maximum s-defective cliques.

Corollary 3.3.3. *The global minimizers of h are all the points p of the form $p = (x^{(C^*)}, y^{(p)})$ where C^* is an s-defective clique of maximum cardinality, and $y^{(p)} \in \{0,1\}^{\overline{E}}$ such that $\mathbf{1}^T y^{(p)} = \min(s, |\overline{E}|)$.*

Proof. Let $p = (x^{(C^*)}, y^{(p)})$ be a local minimizer for h and denote $l = \mathbf{1}^T y^{(p)}$. Then its objective value is, by (31), $h(p) = \alpha |C^*|^{1-k} - \beta l$, which is globally minimized when $|C^*|$ and l are as large as possible under the constraints $l \leq s$ and $l \leq |\overline{E}|$ (because $y^{(p)} \in \{0,1\}^{\overline{E}}$). ☐

3.4 A continuous characterization of maximal s-plexes in k-graphs

3.4.1 Introduction

This result was not present in [9], but almost all the ideas (minus the regularization term ensuring full equivalence between the formulations) were already present in [34] so the above continuous formulation could easily be extended from the s-defective clique case to the s-plex case with just a minor change.

3.4.2 Specific definitions and notations

An s-plex is a subset S of vertices such that each vertex in S is adjacent to at least $|S| - s$ other vertices from S. An s-plex is said to be maximal if it is not contained in any other s-plex, while it is called maximum if it has maximum cardinality. Finding the largest s-plex in a graph is known as the Maximum s-Plex Problem (MsPP), which is a combinatorial problem.

For a given k-graph $G = (V, E)$, denote by $A(\overline{G}) \in \{0,1\}^{|V| \times |\overline{E}|}$ the incidence matrix of its complement graph, defined by :

$$\forall e \in \overline{E},\ \forall i \in V,\ \ \big(A(\overline{G})\big)_{ie} = \begin{cases} 1 \text{ if } i \in e \\ 0 \text{ else} \end{cases} \tag{52}$$

i.e. each row corresponds to a vertex $i \in V$, and each column corresponds to an edge $e \in \overline{E}$, such that, for a given $y \in \{0,1\}^{|\overline{E}|}$ coding for the addition of fake edges, doing the product $A(\overline{G})y$ yields a vector of size $|V|$ giving the number of fake edges each vertex belongs to.

Let $\mathcal{D}_s(G) := \{y \in [0,1]^{\overline{E}} | A(\overline{G})y \leq (s-1)\mathbf{1}\}$ and $\mathcal{P}_s := \Delta \times \mathcal{D}_s(G)$.

The rest of the subsection is dedicated to the derivation of a continuous formulation of the MsPP.

3.4.3 Previous work

In [34], the authors provide an adaptation of their formulation for the MsdCP, which is simply a change in the search space :

$$\max_{\text{s.t. } (x,y) \in \Delta \times \mathcal{D}_s(G)} x^T (A + A(y)) x \tag{53}$$

where $\Delta = \{x \in \mathbb{R}^n | x \geq 0 \text{ and } 1^T x = 1\}$ and $\mathcal{D}_s(G) = \{y \in \{0,1\}^{\overline{E}} | By \leq (s-1)\mathbf{1}\}$ where B is the incidence matrix of graph G (which replace $\mathcal{D}_s(G) = \{y \in \{0,1\}^{\overline{E}} | 1^T y \leq s\}$ in the s-defective clique case). Again, $y \in \{0,1\}^{\overline{E}}$ in the definition of $\mathcal{D}_s(G)$ can be relaxed to $y \in [0,1]^{\overline{E}}$ in order to obtain a fully continuous formulation.

3.4.4 Formulation considered

The same adaptation can be applied to our formulation.

Let $G := (V, E)$ be a k-graph with vertices V and edges E. Let $0 < \alpha \leq \frac{1}{k(k-1)}$ (with strict inequality for $k = 2$) and $\beta > 0$ and consider the following problem :

$$\min_{\text{s.t. } (x,y) \in \mathcal{P}_s} L_{\overline{G}}(x) - L_{G(y)}(x) + \alpha \|x\|_k^k - \beta \|y\|_2^2 \tag{54}$$

where $\mathcal{P}_s = \Delta \times \mathcal{D}_s(G)$ with $\mathcal{D}_s(G) = \{y \in [0,1]^{\overline{E}} | A(\overline{G})y \le (s-1)\mathbf{1}\}$ (which still is a polytope) instead of $\{y \in [0,1]^{\overline{E}} | \mathbf{1}^T y \le s\}$.

We claim that minimizers of this problem are attained at $p = (x^{(C)}, y^{(p)})$, where $s - 1 \ge l = \max(A(\overline{G})y) \in \mathbb{N}$, with C an l-plex in G which is also a maximal clique in $G \cup G(y^{(p)})$, and $y^{(p)} \in \{0,1\}^{\overline{E}}$ such that $y_e^{(p)} = 1$ for every $e \in \binom{C}{k} \cap \overline{E}$ and with $\mathrm{supp}(y^{(p)})$ of maximum cardinality under these constraints, and conversely we also claim that every maximal s-plex in G is associated to such a minimizer.

3.4.5 Correctness of the formulation

The reasoning of the previous part can be almost exactly applied to this new problem with the exception of the application of the theorem given by [10] (*MCP in k-graphs*) at the end of the proof of $(i) \Rightarrow (iii)$ in 3.3.2. As the result is applied to an augmented graph where $y^{(p)}$ cannot add more than $s-1$ edges to each node, the resulting maximal clique in this setting becomes a maximal l-plex instead of a maximal l-defective clique.

3.5 Results for variants of the FW algorithm

3.5.1 Introduction

Now that we had a continuous formulation for the MsdCP in hypergraphs with proved equivalence, we could move to the generalization of various results about the convergence of two algorithms (which are variants of the FW algorithm) solving the above continuous problem. Again, the proofs in [9] (minus a theorem we failed to generalize, which is discussed at the end of the subsection) were easily adapted to this new formulation, and again after suggestions from Mr. Zeffiro in order to lighten the notations when working with tensors, we arrived to the following results and proofs.

In case you do not know what the FW algorithm is, it is an optimization scheme especially suited when the search space is a polytope (as it is in our case). At each step, the algorithm computes the gradient $\nabla f(x_k)$ of the objective function f at the current position x, and move towards the vertex v_k minimizing the dot product $\nabla f(x_k)^T v_k$, where the step size can be chosen by any strategy, as long as it is above 0 and below 1. As the search space is convex and x_k and v_k are both contained within it, the updated position x_{k+1} is necessarily feasible, which eliminates the need for a costly projection step into the feasible domain.

Its pseudo-code is detailed hereafter :

Algorithm 1 FW Algorithm

Initialize x^0 within the feasible polytope $\mathcal{P} = \mathrm{conv}(\mathcal{Q})$ with $|\mathcal{Q}| < \infty$.
for $t = 1, 2, \ldots$ **do**
  Compute the gradient: $\nabla f(x^t)$.
  Find the vertex $v \in \mathcal{Q}$ that maximizes $\nabla f(x^t)^T v$.
  Determine the step size: γ_t.
  Update the solution: $x^{t+1} = x^t + \gamma_t(v - x^t)$.
end for

3.5.2 Specific notations and definitions

In the following, we will denote $\tilde{h} := -h$ to stay in accordance with [9], which work on a maximization problem instead of a minimization problem like here.

Let $\Delta^{(C)} := \{x \in \Delta | x_i = 0 \text{ for all } i \in V \setminus C\}$ be the minimal face of Δ containing $x^{(C)}$ in its relative interior. We recall that in a linear problem a face is the set of all points satisfying the equalities and inequalities constraints of the problem where some inequalities are transformed into equalities.

We define the face of a polytope $\mathcal{Q}$ exposed by a gradient $g \in \mathbb{R}^n$ as

$$\mathcal{F}_e(g) := \arg\max_{w \in \mathcal{Q}} g^T w \tag{55}$$

For every $x \in \Delta$, $i \in [1, \ldots, n]$, the multiplier function $\lambda_i : \Delta \to \mathbb{R}$ for a function f is defined as

$$\lambda_i(x) := \nabla f(x)^T(\mathbf{e}_i - x) \tag{56}$$

or in vector form

$$\lambda(x) := \nabla f(x) - x^T \nabla f(x) \mathbf{1} \tag{57}$$

We recall the FDFW algorithm, applied on a polytope $\mathcal{Q} := \mathrm{conv}(A) \subset \mathbb{R}^n$ with $|A| < +\infty$ with objective function f :

Algorithm 2 FDFW

1: **Initialize:** $w_0 \in \mathcal{Q}$, $k = 0$
2: **if** w_k is stationary **then**
3:     **STOP**
4: **end if**
5: Find $s_k \in \arg\max_{y \in \mathcal{Q}} \nabla f(w_k)^T y$ and $d_k^{\mathcal{FW}} = s_k - w_k$.
6: Find $v_k \in \arg\min_{y \in \mathcal{F}(w_k)} \nabla f(w_k)^T y$ and $d_k^{\mathcal{FD}} = w_k - v_k$.
7: **if** $\nabla f(w_k)^T d_k^{\mathcal{FW}} \geq \nabla f(w_k)^T d_k^{\mathcal{FD}}$ **then**
8:     $d_k = d_k^{\mathcal{FW}}$
9: **else**
10:     $d_k = d_k^{\mathcal{FD}}$
11: **end if**
12: Choose the step size $\alpha_k \in (0, \alpha_k^{\max}]$ using a suitable criterion.
13: Update $w_{k+1} = w_k + \alpha_k d_k$.
14: Set $k = k + 1$ and go to step 2.

At every iteration, the FDFW chooses between the classic FW direction $d_k^{\mathcal{FW}}$ and the in face direction $d_k^{\mathcal{FD}}$. The classic FW direction points toward the vertex maximizing the scalar product with the current gradient or equivalently the vertex maximizing the first order approximation $w \to f(w_k) + \nabla f(w_k)^T w$ of the objective f. The in face direction $d_k^{\mathcal{FD}}$ is always a feasible direction in $\mathcal{F}(w_k)$ from w_k and it points away from the vertex of the face minimizing the first order approximation of the objective.

Remark 3.5.1. *For a more illustrated explanation, take $\mathcal{Q} = \Delta$. The problem can be seen as the attribution of a budget, and we have two possibilities to spend it better :*

- *We take some budget from the other vertices and give it to a profitable one (which corresponds to the classic FW direction $d_k^{\mathcal{FW}}$)*
- *Or we take the budget of a poor vertex and give it to the others (which corresponds to the in face direction $d_k^{\mathcal{FD}}$)*

When $f = \tilde{h}$ and $\mathcal{Q} = \mathcal{P}_s$, it is not difficult to see that the main cost to compute v_k is finding the smallest s components of a vector with size at most $|\overline{E}|$. After the algorithm performs an in face step, we have that the minimal face containing the current iterate either stays the same or its dimension drops by one. The latter case occurs when the method performs a maximal feasible in face step (i.e. a step with $\alpha_k = \alpha_k^{\max}$ and $d_k = d_k^{\mathcal{FD}}$), generating a point on the boundary of the current minimal face. As proved in **Proposition 7.2** of [9], this drop in dimension is what allows the method to quickly identify low dimensional faces containing solutions.

We often require the following lower bound on the stepsizes :

$$\alpha_k \geq \overline{\alpha}_k := \min\left(\alpha_k^{\max}, c\frac{\nabla f(w_k)^T d_k}{\|d_k\|_2^2}\right) \tag{58}$$

Furthermore, for some convergence results we need the following sufficient increase condition (used in the Armijo / backtracking line search) for some constant $\rho > 0$:

$$f(w_k + \alpha_k d_k) - f(w_k) \geq \rho\overline{\alpha}_k \nabla f(w_k)^T d_k \tag{59}$$

As showed for the case $k = 2$ of the MsdCP in [9], the convergence of the previous optimization scheme applied to our objective can be slow and inefficient. Since y is tied to x, it is not possible to efficiently change the regularization parameters to speed up convergence (either the algorithm ignores x if the penalty coefficient on the y variable is large or it ignores y if this coefficient is small). This motivated the authors to introduce a tailored FW variant, named FWdc, exploiting the cross product structure of the search space by splitting the update rules, which is recalled hereafter :

Algorithm 3 FWdc

1: **Initialize:** $z_0 = (x_0, y_0) \in \Delta \times \mathcal{D}_s(G)$, $k = 0$
2: **if** z_k is stationary **then**
3:     **STOP**
4: **end if**
5: Compute x_{k+1} by applying one step of the previous algorithm with $w_0 = x_k$ and $f(w) = \tilde{h}(w, y_k)$.
6: Find $y_{k+1} \in \arg\max_{y \in \mathcal{D}_s(G)} \nabla_y h_G(x_{k+1}, y_k)^T y$.
7: Set $k = k + 1$ and go to step 2.

At every iteration the method alternates an FDFW step on the x variable with a full FW step on the y variable so that y_k always belongs to $\{0,1\}^{\overline{E}}$.

The rest of the subsection is dedicated to the generalization of the identification and convergence results for the FDFW and FWdc variants that were found in [9], applied to the objective $\tilde{h}$.

3.5.3 Results

For any of the two considered problems, thanks to the previous parts we can express any maximizer p of $\tilde{h}$ as $p = (x^{(C)}, y^{(p)})$ where $y^{(p)} \in \{0,1\}^{\overline{E}}$ such that $\mathbf{1}^T y^{(p)} \leq s$ and C is a maximal clique of $G \cup G(y^{(p)})$.

We now prove that the face of $\mathcal{P}_s$ exposed by the gradient in p a maximizer is simply the product between Δ and the singleton $\{y^{(p)}\}$. This property, sometimes called strict complementarity, is of key importance to prove identification results for variants of the FW algorithm (see [8], [7] and [13], and the discussion of external regularity in **subsection 5.3** of [6]).

Lemma 3.5.2. *Let $p := (x^{(C)}, y^{(p)})$ a strict minimizer. Then the face exposed by $\nabla\tilde{h}(p)$ coincides with the minimal face $\mathcal{F}(p)$ of $\mathcal{P}_s$ containing p :*

$$\mathcal{F}_e(\nabla\tilde{h}(p)) = \mathcal{F}(p) = \Delta^{(C)} \times \left\{y^{(p)}\right\} \tag{60}$$

Proof. To start with, the second equality follows from the fact that $y^{(p)}$ is a vertex of $\mathcal{D}_s(G)$ and that $\Delta^{(C)}$ is the minimal face of Δ containing $x^{(C)}$. The first equality is then equivalent to proving that for every vertex $a := (a_x, a_y)$ of $\mathcal{P}_s$ with $a \in \mathcal{P}_s \setminus \mathcal{F}(p)$ we have $\lambda_a(p) < 0$. Given that stationarity conditions must hold in Δ and $\mathcal{D}_s(G)$ separately, $\lambda_a(p) < 0$ iff

$$\lambda_a^x(p) = \nabla_x \tilde{h}(p)^T (a_x - x^{(C)}) \leq 0 \tag{61}$$

$$\lambda_a^y(p) = \nabla_y \tilde{h}(p)^T (a_y - y^{(p)}) \leq 0 \tag{62}$$

and at least one of these relations must be strict. Since a is a vertex of $\mathcal{P}_s$, $a_x = \mathbf{e}_l$ with $l \in [1, \dots, n]$ and $\{0,1\}^{\overline{E}}$, while $a \notin \mathcal{F}(p)$ implies $l \notin C$ or $a_y \neq y^{(p)}$. We have

$$\nabla_x \tilde{h}(p)^T x^{(C)} = \sum_{j=1}^{n} x_j^{(C)} \frac{\partial \tilde{h}}{\partial x_j}(p) = -\sum_{j \in C} x_j^{(C)} \left[\sum_{e \in \overline{E} \setminus E(y)} \mathbf{1}_{j \in e} \prod_{i \in e \setminus \{j\}} x_i^{(C)} + \alpha k \left(x_j^{(C)} \right)^{k-1} \right] \tag{63}$$

$$= -\sum_{j \in C} \left[\sum_{e \in \overline{E} \setminus E(y)} \mathbf{1}_{j \in e} \prod_{i \in e} x_i^{(C)} + \alpha k \left(x_j^{(C)} \right)^{k} \right] \tag{64}$$

$$= -\alpha k \|x^{(C)}\|_k^k \tag{65}$$

and

$$\nabla_x \tilde{h}(p)^T a_x = \frac{\partial \tilde{h}}{\partial x_l}(p) = -\sum_{e \in \overline{E} \setminus E(y)} \mathbf{1}_{l \in e} \prod_{i \in e \setminus \{l\}} x_i^{(C)} - \alpha k \left(x_l^{(C)} \right)^{k-1} = -\alpha k \left(x_l^{(C)} \right)^{k-1} \tag{66}$$

because the clique is maximal. Combining the two, we obtain

$$\lambda_a^x(p) = \nabla_x \tilde{h}(p)^T (a_x - x^{(C)}) = \alpha k \left(x_l^{(C)} \right)^{k-1} - \alpha k \|x^{(C)}\|_k^k \tag{67}$$

which is equal to 0 if $l \in C$ and is strictly negative otherwise. This proves that (61) holds with strict inequality if $l \notin C$ or else with equality if $l \in C$.

In a similar vein we proceed with (62). If $a_y = y^{(p)}$, then (62) holds with equality, but then $l \in V \setminus C$, and we are done. So assume $a_y \neq y^{(p)}$, and consider the supports $S_a = \left\{ e \in \overline{E} | (a_y)_e = 1 \right\}$ and $S_y = \left\{ e \in \overline{E} | y_e^{(p)} = 1 \right\}$.

Necessarily, $S_y \setminus S_a \neq \emptyset$. Indeed, whether we are in the case of the MsdCP or in the case of the MsPP, both a_y and $y^{(p)}$ are elements of the same polytope i.e. they satisfy the same constraints, and, because p is a minimizer, $y^{(p)}$ has a support of maximal cardinality under these constraints, so if $S_y \subset S_a$ then $|S_y| = |S_a|$, but because $a_y \neq y^{(p)}$ then $S_y \not\subset S_a$ i.e. $S_y \setminus S_a \neq \emptyset$.

Then, for every e in S_y we have

$$\frac{\partial \tilde{h}}{\partial y_e}(p) = \prod_{i \in e} x_i^{(C)} + 2\beta y_e^{(p)} \geq 2\beta > 0 \tag{68}$$

while for every e in $S_a \setminus S_y$ we have

$$\frac{\partial \tilde{h}}{\partial y_e}(p) = 0 \tag{69}$$

because $y_e^{(p)} = 0$ by definition of S_y and $e \in \overline{E} \setminus E(y^{(p)})$. Then

$$\lambda_a^y(p) = \nabla_y \tilde{h}(p)^T (a_y - y^{(p)}) = \sum_{e \in S_a} \frac{\partial \tilde{h}}{\partial y_e}(p) - \sum_{e \in S_y} \frac{\partial \tilde{h}}{\partial y_e}(p) \tag{70}$$

$$= \sum_{e \in S_a \setminus S_y} \frac{\partial \tilde{h}}{\partial y_e}(p) - \sum_{e \in S_y \setminus S_a} \frac{\partial \tilde{h}}{\partial y_e}(p) \tag{71}$$

$$= -\sum_{e \in S_y \setminus S_a} \frac{\partial \tilde{h}}{\partial y_e}(p) < 0 \tag{72}$$

which proves that for any minimizer p, (61) and (62) both holds with one of the two strictly. $\square$

This result allows us to prove the following local convergence and identification result for the FDFW applied to our maximal s-defective clique formulation.

Theorem 3.5.3 (FDFW local identification and convergence)**.** *Let $p := (x^{(C)}, y^{(p)})$ be a strict minimizer, let z_k be a sequence generated by the FDFW. Then under (58) there exists a neighborhood $U(p)$ of p such that if $K := \min\{k \in \mathbb{N} | z_k \in U(p)\}$ we have the following properties :*

(a) if $\tilde{h}(z_k)$ is monotonically increasing, then supp$(z_k) = C$ and $y_k = y^{(p)}$ for every $k \geq K + dim\ \mathcal{F}(w_k)$

(b) if (59) also holds, then $z_k \to p$.

Proof. $\tilde{h}(\cdot, y^{(p)})$ is strongly concave in $\Delta^{(C)}$ (see (51) bearing in mind $\tilde{h} = -h$), and by 3.5.2, we have that $\mathcal{F}_e(\nabla \tilde{h}(p)) = \mathcal{F}(p) = \Delta^{(C)} \times \{y^{(p)}\}$. We have all the necessary assumptions to apply **Lemma 7.4** of [9], which we recall hereafter :

> ***Lemma :*** *Let p be a local maximizer for f restricted to $\mathcal{Q}$. Assume that (58) holds and that f is strongly concave in $\mathcal{F}_e(\nabla f(p))$. Then, for a neighborhood $U(p)$ of p, if $w_0 \in U(p)$,*
>
> *(a) if $\{f(w_k)\}$ is increasing, there exists $k \in [0, \dots, dim(\mathcal{F}(w_0))]$ such that $w_{k+i} \in \mathcal{F}_e(\nabla f(p))$ for every $i \geq 0$*
>
> *(b) if in addition (59) holds, then $\{w_{k+i}\}_{i \geq 0}$ converges to p.*

□

As a corollary, we have the following global convergence result under the mild hypothesis that the set of limit points contains no saddle points.

Corollary 3.5.4 (FDFW global convergence)**.** *Let $\{z_k\}$ be a sequence generated by the FDFW, and assume that there are no saddle points in the limit set of $\{z_k\}$. Then under the conditions (58) and (59) on the step size we have $z_k \to p := (x^{(C)}, y^{(p)})$ with p a strict minimizer such that supp$(x_k) \subset C$ and $y_k = y^{(p)}$ for k large enough.*

Proof. As for for 3.5.3, we have all the necessary assumptions to apply **Corollary 7.5** of [9], which is recalled hereafter :

> ***Corollary :*** *Let $\{w_k\}$ be a sequence generated by the FDFW algorithm. Assume that there are no saddle points in the limit set of $\{w_k\}$ and that for every local maximizer p the objective f is strongly concave in $\mathcal{F}_e(\nabla f(p))$. Then under the conditions (58) and (59) on the stepsize, we have $w_k \to p$ with p a local maximizer satisfying $w_k \in \mathcal{F}_e(\nabla f(p))$ for k large enought.*

□

As recalled in the introduction of this subsection, the FDFW is not adapted to the structure of our problem, which motivated the authors of [9] to separate the update rules for the two variables into a variant of the FW algorithm they named FWdc. The next proposition proves that in the FWdc the sequence $\{y_k\}$ is ultimately constant, the rest of the iterations then correspond to an application of the FDFW on the graph $G \cup G(\overline{y})$ where $\overline{y}$ is the final value for y, which allows to obtain convergence results by applying the general properties of the FDFW proved before to the x component.

Proposition 3.5.5. *In the FWdc variant, if $\tilde{h}(z_k)$ is increasing at each separate update of x_k and y_k, then $\{y_k\}$ can change at most $l + \frac{|\overline{E}| + \alpha\left(1 - |C^*|^{1-k}\right)}{\beta}$ times, with C^* a maximum l-defective clique if we consider the MsdCP and a maximum l-plex if we consider the MspP.*

Proof. Assume that for a step of the algorithm, y_k and y_{k+1} are distinct vertices of $\mathcal{D}_s$. Then

$$\begin{aligned}
\tilde{h}(z_{k+1}) - \tilde{h}(z_k) &= \tilde{h}(x_{k+1}, y_{k+1}) - \tilde{h}(x_k, y_k) && (73)\\
&= \tilde{h}(x_{k+1}, y_{k+1}) - \tilde{h}(x_{k+1}, y_k) + \tilde{h}(x_{k+1}, y_k) - \tilde{h}(x_k, y_k) && (74)\\
&\geq \tilde{h}(x_{k+1}, y_{k+1}) - \tilde{h}(x_{k+1}, y_k) && (75)\\
&\geq \nabla_y \tilde{h}(x_{k+1}, y_k)^T (y_{k+1} - y_k) + \beta \|y_{k+1} - y_k\|_2^2 && (76)\\
&\geq \beta > 0 && (77)
\end{aligned}$$

where we used the hypothesis that $\tilde{h}(z_k)$ is increasing at each separate update of x_k and y_k in (75), the 2β-strong convexity of $\tilde{h}(x, \cdot)$ in (76), and $y_{k+1} \in \arg\max_{y \in \mathcal{D}'_s} \nabla_y \tilde{h}(z_k)^T y$ and the fact that the distance between vertices of $\mathcal{D}_s$ is at least 1 in the last line.

By summing inequality (77), we get

$$\tilde{h}(z_N) - \tilde{h}(z_0) = \sum_{i=0}^{N-1} \left[\tilde{h}(z_{k+1}) - \tilde{h}(z_k) \right] \geq changes(N)\beta \tag{78}$$

where $changes(N)$ is the number of times y_k changes between index 0 and index N. But using $-\tilde{h}(z_0) = h(z_0) \leq |\overline{E}| + \alpha$ and 3.3.3, we have that for all $N \in \mathbb{N}$

$$\tilde{h}(z_N) - \tilde{h}(z_0) \leq \max_{z \in \mathcal{P}_s} \tilde{h}(z) - \tilde{h}(z_0) \leq \max_{z \in \mathcal{P}_s} \tilde{h}(z) + |\overline{E}| + \alpha \leq \beta l + |\overline{E}| + \alpha \left(1 - |C^*|^{1-k}\right) \tag{79}$$

Thus $\{changes(N)\}_N$ is bounded, so because it is an increasing integer sequence it necessary becomes stationary after a certain rank. By the previous inequality, we can now bound the number of times y_k can change by

$$changes(+\infty) \leq l + \frac{|\overline{E}| + \alpha \left(1 - |C^*|^{1-k}\right)}{\beta} \tag{80}$$

□

Remark 3.5.6. *In a more prosaic way, because the FWdc updates y by the rule*

$$y_{k+1} \in arg\ max_{y \in \mathcal{D}_s(G)} \nabla_y h_G(x_k, y_k)^T y \tag{81}$$

i.e. it searches for the s possible fake edges of maximum derivative values, which are equal to

$$\frac{\partial \tilde{h}}{\partial y_e}(x, y) = \prod_{i \in e} x_i + 2\beta y_e \tag{82}$$

the parameter β can be thought of as an inertia : for a possible fake edge to become selected at this step, its product needs to overcome the gradient value of one already selected fake edge, which benefits from a bonus modulated by β. Thus the higher β is and the more difficult it becomes for y to change.

The following theorem allows to explicitly bound how close the sequence $\{x_k\}$ generated by the FWdc must be to $x^{(C)}$ for the identification to happen.

Proposition 3.5.7. *Let $\{z_k\}$ be a sequence generated by the FWdc, $\overline{y} \in \{0, 1\}^{\overline{E}} \cap \mathcal{D}_s$ (where $\mathcal{D}_s$ is defined in the sense of one of the two considered problems), and C be a clique in $G \cup G(\overline{y})$; let $\delta_{\max}$ be the maximum eigenvalue of the adjacency tensor $\mathcal{A}(\overline{G \cup G(\overline{y})})$.*

For a node $j \in V \setminus C$, we define

$$E^C(j) := \left| \left\{ e \in \overline{E} \setminus E(\overline{y}) | j \in e,\ e \in \{j\} \times \binom{C}{k-1} \right\} \right| \tag{83}$$

the number of missing edges containing node j and $k-1$ nodes of the clique.

We also define

$$m(C, G \cup G(\overline{y})) := \min_{v \in V \setminus C} E^C(v) \tag{84}$$

the minimum number of edges needed to increase by 1 the size of the clique, and

$$m_\alpha(C, G \cup G(\overline{y})) := m(C, G \cup G(\overline{y})) - \alpha k \tag{85}$$

Let K be a fixed index in $\mathbb{N}$ and I^c be the components of supp(x_K) with index not in C, and let $L := \frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha$. Assume that $y_{K+j} = \overline{y}$ is constant for $0 \leq j \leq |I^c|$, that (58) hold for $c := \frac{1}{L}$, and that

$$\|x_K - x^{(C)}\|_1 \leq \frac{m_\alpha(C, G \cup G(\overline{y}))}{m_\alpha(C, G \cup G(\overline{y})) + 2|C|^{k-1}\left(\frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha\right)} \tag{86}$$

Then supp$(x_{K+|I^c|}) = C$.

Proof. Since y_k does not change for $k \in [K, K + |I^c|]$, the FWdc corresponds to an application of the FDFW to the simplex Δ on the variable x. Let

$$\lambda_{\min} := \min_{i \in V \setminus C} -\lambda_i(x^{(C)}) \tag{87}$$

be the smallest negative multiplier with corresponding index not in C. Let L' be a Lipschitz constant for $\nabla_x h(x, y)$ with respect to the variable x. By Theorem 3.3 of [8], if

$$\|x_k - x^{(C)}\|_1 < \frac{\lambda_{\min}}{\lambda_{\min} + 2L'} \tag{88}$$

we have the desired identification result.

We now prove that we can take $L' = L$. In the following, we will abbreviate $L_{\overline{G \cup G(\overline{y})}} := L$ and $\mathcal{A}(\overline{G \cup G(\overline{y})}) := \mathcal{A}$.

$$\|\nabla_x \tilde{h}(x', \overline{y}) - \nabla_x \tilde{h}(x, \overline{y})\|_2 = \|\nabla L(x') - \nabla L(x) + k\alpha(x'^{[k-1]} - x^{[k-1]})\|_2 \tag{89}$$

$$\leq \|\nabla L(x') - \nabla L(x)\|_2 + k\alpha\|(x'^{[k-1]} - x^{[k-1]}\|_2 \tag{90}$$

Then by the mean value theorem we have that $\|x'^{[k-1]} - x^{[k-1]}\|_2 \leq (k-1)\|x' - x\|_2$.

About the first term, denoting $\lambda_{\max}(u)$ the maximum eigenvalue of the Hessian matrix at a point u, we have :

$$\lambda_{\max}(u) = \max_{\|v\|_2=1} |v^T \nabla^2 L(u) v| \tag{91}$$

$$= \frac{1}{(k-2)!} \max_{\|v\|_2=1} |\mathcal{A}u^{k-2}v^2| \tag{92}$$

$$\leq \frac{1}{(k-2)!} \max_{\|v\|_2=1} \frac{|\mathcal{A}u^{k-2}v^2|}{\|u\|_2^{k-2}} \text{ because } \|u\|_2 \leq \|u\|_1 \leq 1 \tag{93}$$

$$\leq \frac{1}{(k-2)!} \max_{\|v_i\|_2=1,\ 1 \leq i \leq k} |\mathcal{A}v_1 \dots v_k| \tag{94}$$

$$= \frac{1}{(k-2)!} \max_{\|v\|_2=1} |\mathcal{A}v^k| = \frac{1}{(k-2)!}\delta_{\max} \tag{95}$$

where we used **(1.2)** and **(1.3)** of [23] for the final equality.

This means that all eigenvalues of the Hessian are upper bounded by $\frac{1}{(k-2)!}\delta_{\max}$, thus

$$\|\nabla L(x') - \nabla L(x)\|_2 \leq \frac{1}{(k-2)!}\delta_{\max}\|x' - x\|_2 \tag{96}$$

Combining the two Lipschitz constants found, we get

$$\|\nabla_x \tilde{h}(x', \overline{y}) - \nabla_x \tilde{h}(x, \overline{y})\|_2 \leq \left(\frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha\right)\|x' - x\|_2 \tag{97}$$

About the multipliers, for $j \in V \setminus C$ we have the equality

$$-\lambda_j(x^{(C)}) = \nabla_x \tilde{h}(x^{(C)}, \overline{y})^T (x^{(C)} - \mathbf{e}_j) \tag{98}$$

$$= \sum_{e \in \overline{E} \setminus E(\overline{y})} \mathbf{1}_{j \in e} \prod_{i \in e \setminus \{j\}} x_i^{(C)} - \frac{1}{|C|}\sum_{l \in C}\left[\sum_{e \in \overline{E} \setminus E(\overline{y})} \mathbf{1}_{l \in e} \prod_{i \in e \setminus \{l\}} x_i^{(C)} + \alpha k \left(\frac{1}{|C|}\right)^{k-1}\right] \tag{99}$$

$$= \frac{|E^C(j)| - \alpha k}{|C|^{k-1}} \tag{100}$$

Thus

$$\lambda_{\min} = \min_{\in V \setminus C} -\lambda_i(x^{(C)}) = \min_{i \in V \setminus C} \frac{|E^C(i)| - \alpha k}{|C|^{k-1}} \tag{101}$$

$$= \frac{m(C, G \cup G(\overline{y})) - \alpha k}{|C|^{k-1}} \tag{102}$$

$$= \frac{m_\alpha(C, G \cup G(\overline{y}))}{|C|^{k-1}} \tag{103}$$

Finally we have

$$\frac{\lambda_{\min}}{\lambda_{\min} + 2L} = \frac{m_\alpha(C, G \cup G(\overline{y}))}{m_\alpha(C, G \cup G(\overline{y})) + 2|C|^{k-1}\left(\frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha\right)} \tag{104}$$

□

Remark 3.5.8. *This result is very similar to what was found in [9] for the case $k = 2$: we have $L := \frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha$ where $0 < \alpha \leq \frac{1}{k(k-1)}$ with strict inequality for $k = 2$ instead of $L := 2\delta + \alpha$ where $0 < \alpha < 2$ and*

$$\frac{m_\alpha(C, G \cup G(\overline{y}))}{m_\alpha(C, G \cup G(\overline{y})) + 2|C|^{k-1}\left(\frac{1}{(k-2)!}\delta_{\max} + k(k-1)\alpha\right)} \tag{105}$$

with $m_\alpha(C, G \cup G(\overline{y})) := m(C, G \cup G(\overline{y})) - \alpha k$ instead of

$$\frac{m_\alpha(C, G \cup G(\overline{y}))}{m_\alpha(C, G \cup G(\overline{y})) + 2|C|\delta_{\max} + |C|\alpha} \tag{106}$$

with $m_\alpha(C, G \cup G(\overline{y})) := m(C, G \cup G(\overline{y})) - 1 + \frac{\alpha}{2}$.

3.6 A theorem we failed to generalize

We failed to generalize a last theorem :

> **Theorem 5.2** Let $\{z_k\}$ be a sequence generated by the FDFW, with regularization coefficient $\alpha = 1$. If conditions (58) and (59) hold on the step sizes, then $\{z_k\}$ converges to a stationary point and identifies its support infinite time.

For that, we need both that there is a finite number of stationary points in each face of the polytope, and that these stationary points satisfy strict complementarity. I tried to adapt the proofs for both these points to the case $k > 2$ with no success. After a discussion with Mr. Zeffiro, he sent me these notes containing an alternative and better proof for the case $k = 2$ (the condition $\alpha = 1$ was relaxed to $\alpha \in \mathbb{Q}$) that I couldn't generalize either.

Both of the proofs for $k = 2$ rely on the gradient of the objective function : as it is quadratic, when derived only one term in x remains in the product, and various tricks can be applied on this sole x term. Meanwhile for $k > 2$, $k-1$ terms remain in the derived product, forbidding the use of the same tricks.

3.7 Numerical results

3.7.1 Introduction

In this subsection we report on preliminary numerical comparison of the methods, obtained with the implementation available here. In the test, to satisfy the condition $0 < \alpha \leq \frac{1}{k(k-1)}$ with strict inequality when $k = 2$, we set $\alpha = 0.4$ for $k = 2$ and $\alpha = \frac{1}{k(k-1)}$ in the other cases, and $\beta = \frac{1}{2n^k}$. A simple reason for this choice is that, as remarked in 3.5.6, β acts as a bonus for selected fake edges to keep being selected, thus it needs to be small enough that the product of an unselected edge can overcome it.

More formally, if $x_k = x^{(C)}$ with C an s-defective clique and $(y_k)_{ij} = 0$ for $e \in \binom{C}{k}$, we want to ensure that the FW vertex $s_k = (x^{(s_k)}, y^{(s_k)})$ is such that $y_{ij}^{(s_k)} = 1$. Now for $e' \notin \binom{C}{k}$ and assuming $|C| < n$ (otherwise $C = V$ and the problem is trivial), we have

$$\frac{\partial \tilde{h}}{\partial y_e}(x_k, y_k) = \left(\frac{1}{|C|}\right)^k > \left(\frac{1}{n}\right)^k = 2\beta = \frac{\partial \tilde{h}}{\partial y_{e'}}(x_k, y_k) \tag{107}$$

From this it is then immediate to conclude that e must be in the support of $y^{(s_k)}$.

3.7.2 Implementation

I implemented the two algorithms (FDFW and FWdc) both in a naive way and integrating the Short Steps Chain (SSC) procedure detailed in [31], the latter version being used in the tests. This procedure aims at improving existing convergence rates for FW variants and improves the execution speed by skipping unnecessary gradient computations. To explain briefly what it is, the procedure is integrated into another first-order method (in our case the two FW variants), as in the following pseudo-code :

Algorithm 4 First-order method with SSC

Initialization: $x_0 \in \Omega$, $k = 0$.

```
1: while x_k is not stationary: do
2:     g = −∇f(x_k)
3:     x_{k+1} = SSC(x_k, g)
4:     k = k + 1
5: end while
```

Algorithm 5 SSC($\bar{x}, g$)

Initialization: $y_0 = \bar{x}$, $j = 0$.

Phase I

1: Select $d_j \in A(y_j, g), \alpha_{\max}^{(j)} \in \alpha_{\max}(y_j, d_j)$
2: **if** $d_j = 0$ **then**
3:     **return** y_j
4: **end if**

Phase II

1: Compute β_j with (109)
2: Let $\alpha_j = \min(\alpha_{\max}^{(j)}, \beta_j)$
3: $y_{j+1} = y_j + \alpha_j d_j$
4: **if** $\alpha_j = \beta_j$ **then**
5:     **return** y_{j+1}
6: **end if**
7: $j = j + 1$, go to Step 1.

Considering a classical first-order minimization scheme as in **Algorithm** 4, the SSC procedure replaces the step update part by **Algorithm** 5. Denoting $\bar{B}_r(c)$ the ball of center c and radius r and L the Lipschitz constant of the objective function, the auxiliary step size β_j is defined as the maximal feasible step size for the trust region

$$\Omega_j = \bar{B}_{\|g\|/2L}(\bar{x} + \frac{g}{2L}) \cap \bar{B}_{\langle g, d_j \rangle / L}(\bar{x}), \tag{108}$$

when $y_j \in \Omega_j$, otherwise the method stops returning y_j. Summarizing,

$$\beta_j = \begin{cases} 0 & \text{if } y_j \notin \Omega_j, \\ \beta_{\max}(\Omega_j, y_j, d_j) & \text{if } y_j \in \Omega_j, \end{cases} \tag{109}$$

where $\beta_{\max}(\Omega_j, y_j, d_j)$ is the maximal feasible step size in the direction d_j starting from y_j with respect to Ω_j.

The Lipschitz constant can be iteratively approximated in the following way : at the step k, we start with an estimate $\widetilde{L} = L_k$ of the Lipschitz constant ; then, we compute x_k^+ with the procedure SSC$(x_k, -\nabla f(x_k))$, and repeat setting $\widetilde{L} := 2\widetilde{L}$ until

$$f(x_k) - f(x_k^+) \geq \frac{1}{2} \nabla f(x_k)^\top (x_k - x_k^+) \tag{110}$$

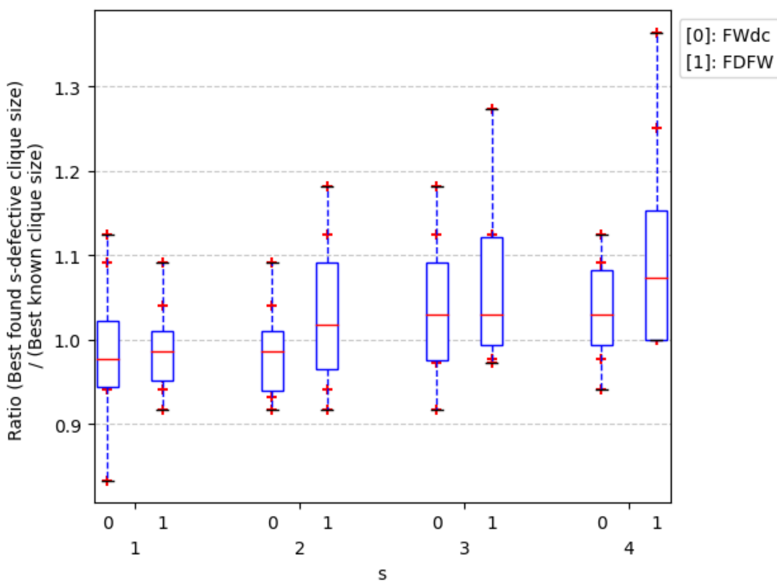
holds. When this condition is met, we set $x_{k+1} = x_k^+$ and $L_{k+1} = \widetilde{L}$.

In the code, common parts shared between a naive implementation of an algorithm and its SSC version has been modularized as external functions that are used by both implementations. To simplify the handling of parameters, lambda functions were heavily employed to reduce the number of input variables at each step of the problem. The variable x is implemented as a Numpy ndarray with a length of $|V|$, while y is implemented as a symmetrical Numpy ndarray of shape $|V|^k$, treated as a symmetrical tensor. Both the naive implementations of these algorithms employ an Armijo line search technique for the selection of the step size. In the context of the FWdc algorithm, the FDFW component applied solely to the x variable has been encapsulated as a standalone function (used with $max\_iteration = 1$ within the FWdc) with the goal of comparing the two, because for $k > 2$ the FDFW becomes excessively time-consuming due to the high dimensionality of the y component. Finally, as we proved that y becomes stationary after a certain iteration and because its update is the most costly part of the algorithm (due to its dimensionality), we investigated an heuristic skipping the updates of y after a certain predetermined number of iterations without changes. It did not prove satisfying as no value of this parameter kept the same solution quality, so it has been left in the code but is not used in practice.

3.7.3 Results

By the end of my internship, we tested the two algorithms on several instances of graphs drawn from the DIMACS implementation challenge (see [16]) for the value $k = 2$, in order to replicate the results found in [9]. The test setting is the following : as for the tests done in [9], values for s (the allowed number of missing edges in the clique) vary from $s = 1$ to $s = 4$; x is randomly initialized on the simplex using a Dirichlet distribution and all the values of y are initialized as 0 ; each algorithm had up to 10 minutes (if a run was not finished by then we simply wait for its end before breaking the loop) and up to 100 random restarts on each graph ; the graphs used for the preliminary tests were *"C125.9"*, *"C250.9"*, *"brock200_2"*, *"brock200_4"*, *"keller4"*, *"p_hat300-1"*, *"p_hat300-2"* and *"p_hat300-3"*. Every time, only the best value was kept. The results are plotted in Figure 1.

Figure 1: Results by varying value of s



On the x-axis are the values of s used, on the y-axis are the ratios (Best found s-defective clique size) / (Best known clique size) for each graph, where the best known clique sizes are taken from the DIMACS dataset, which keeps track of them. Remind that, as we are searching for s-defective cliques, this ratio can be higher than 1. The red line represents the median value in each series, and the boxes extend from the 25th percentile value q_1 to the 75th percentile value q_3. The whiskers cover all the other values in a range of $[q_1 - w(q_3 - q_1), q_3 + w(q_3 - q_1)]$, with the coefficient w equal to 2.7 times the standard deviation of the values.

As can be seen, contrary to the results in [9], with my implementation the FDFW performs better than the FWdc, whereas in the previous implementation the two algorithms behaved similarly. This prompted us to recheck the implementations and, because we couldn't find any flaw that could explain such a gap in performance, we compared the results obtained with my implementation of the FWdc with the results obtained with theirs, using the supplementary material found in [9]. Table 1 sums up the comparison for $s = 4$, on the left are the figures obtained with my implementation, on the right are the figures obtained with the implementation in [9].

The results are strikingly similar, modulo a bit of randomness due to the number of samples (for large graphs each run is costly in terms of time so the 10 minutes of allowed time only allow for a few samples). Thus it seems that it is not my implmentation of the FWdc that performs worse than it should, but that my implementation of the FDFW performs better than the one in [9].

Table 1: Comparison of the implementations

Graph	Mean	Std	Max
brock200_2	9.6 vs 9.5	0.8 vs 0.86	11 vs 12
brock200_4	14 vs 14.1	1.06 vs 1.14	16 vs 17
brock400_2	20.6 vs 21.8	1.35 vs 1.16	23 vs 25
brock400_4	21 vs 21.6	1.34 vs 1.38	25 vs 25
hamming10-4	32.35 vs 32.9	2.5 vs 2.32	35 vs 37
keller4	10.5 vs 10.6	0.72 vs 0.82	12 vs 13
keller5	20 vs 21.5	1.4 vs 1.55	23 vs 27
MANN_a27	121 vs 121.1	0.4 vs 0.38	122 vs 122
MANN_a45	390 vs 334	0 vs 0.22	390 vs 335
p_hat300-1	7.9 vs 7.5	0.89 vs 0.85	10 vs 9
p_hat300-2	23 vs 22.2	1.46 vs 1.27	26 vs 26
p_hat300-3	32 vs 32.2	1.14 vs 1.19	34 vs 36
p_hat700-1	8.97 vs 8.6	0.875 vs 0.77	11 vs 10
p_hat700-2	38.97 vs 40	1.494 vs 2	42 vs 44
p_hat700-3	55.5 vs 57.9	2.69 vs 1.94	60 vs 62

3.7.4 Future work

At the end of my internship we were investigating the source of the difference between my FDFW implementation and the one in [9]. Apart from that, my implementations were time-consuming and could not be used for $k = 4$ and higher values, so a first goal would be to optimize the code. Then, extensive comparisons between the FWdc and the FDFW (as was done in [9] for the case $k = 2$) also need to be done for $k = 3$ and higher values of k if possible. After that, as the formulation we crafted is applied to unweighted graphs, a direction for future research would be to extend this work to weighted hypergraphs.

4 Study of the regularizer in the continuous formulation of the classical MCP in k-uniform hypergraphs

4.1 Introduction

The second part of my internship was dedicated to the search for a generalization of the result found in [15], which studies the choice of the regularizer function in the continuous formulation of the MCP (which originally is simply $\frac{1}{2}\|\cdot\|_2^2$). As shown in the paper, all regularizers are not equivalent and some yield a better solution quality. The result is briefly summarized hereafter :

Consider the following problem :

$$\max_{\text{s.t. } x\in\Delta} x^T Ax + \phi(x) \tag{111}$$

where ϕ satisfies the following conditions :

- $\nabla^2\phi(x) > 0$ i.e. ϕ is strictly convex
- $\|\nabla^2\phi(x)\|_2 < 2$
- $\phi(x)$ is symmetric / invariant by permutation of x

The global and local maximizers of the above problem are strictly characteristic vectors of respectively maximum and maximal cliques.

Examples of such regularizers are :

- $\phi(x) = \frac{1}{2}\|x\|_2^2$
- $\alpha_1\|x + \epsilon 1\|_p^p$ with $\epsilon \geq 0$, $p \geq 2$ and $0 \leq \alpha_1 \leq \frac{2}{p(p-1)(1+\epsilon)^{p-2}}$
- $\phi(x) = \alpha_2 \sum_{i=1}^n (e^{-\beta x_i} - 1)$, with $\beta > 0$ and $0 \leq \alpha_2 \leq \frac{2}{\beta^2}$ (approximation of $-\alpha_2\|x\|_0$)

Like I did for the first part of the internship, I wanted to adapt the proofs in [15], and managed to do so for most of it. Initially I tried to prove a result with weaker hypotheses (the condition **(C3)** in the following was replaced by $\nabla^2\Phi(x) < cst$ in order to have something similar to [15]) but this did not prove successful, so after a discussion with Mr. Zeffiro, we decided to use the conditions that are detailed after. He sent me notes for a simpler and working proof of equivalence between the discrete and continuous formulations which can be accessed here. The following is the fully detailed and rigorous version of these notes.

After some definitions and preliminaries in subsections 4.2 and 4.1, a general regularized formulation of the MCP for hypergraphs is developed in subsection 4.3 and conditions under which the global / local maximizers of the regularized program are in one-one correspondence with the maximum / maximal cliques in G are provided. In subsection 4.5, some particular subsets of these regularizers are studied and some adaptations to the examples given in [15] are provided.

4.2 Definitions and notations

We define the following set :

$$\Delta^0 := \underset{C \text{ clique}}{\cup} \Delta^{(C)} = \{x \in \Delta | \text{supp}(x) \text{ is a clique}\} \tag{112}$$

We define the set of permutations of x as

$$\mathcal{P}(x) := \left\{\overline{x} | \exists \sigma \in \mathcal{S}_n,\, \overline{x}_{\sigma(i)} = x_i \ \forall i \in [1, \ldots, n]\right\} \tag{113}$$

For a function Φ and a set S, we denote by Φ_S the restriction of Φ to S.
We denote by Δ^C the elements of Δ with support equal to C.

4.3 Formulation considered

Let $G := (V, E)$ be a k-graph with vertices V and edges E and consider the following problem :

$$\min_{\text{s.t. } x \in \Delta} L_{\overline{G}}(x) + \Phi(x) \tag{114}$$

where Φ satisfies the three following assumptions :

(C1) $\Phi \in \mathcal{C}^2(\mathbb{R}^n)$ and for all face S of Δ, $\nabla^2\Phi_S(x) > 0$ for all $x \in \Delta^{(S)}$ i.e. the restriction of Φ to any face of Δ is strictly convex

(C2) $\Phi(\overline{x}) = \Phi(x)$ for all $\overline{x} \in \mathcal{P}(x)$ i.e. Φ is symmetric / permutation invariant

(C3) $(\mathbf{e}_i - \mathbf{e}_j)^T \nabla^2\Phi(x)(\mathbf{e}_i - \mathbf{e}_j) < x_i^{k-2} + x_j^{k-2}$ for every $x \in \Delta$, $i, j \in \text{supp}(x)$ with $i \neq j$

In the following, we will denote $h(x) := L_{\overline{G}}(x) + \Phi(x)$. We claim that minimizers of this problem are strict and in one-one correspondence with maximal cliques in G.

Remark 4.3.1. *For $k = 2$, we recover a condition **(C3)** similar to condition **(C3)** from [15], which was $\nabla^2\Phi(x) < 2$.*

4.4 Proof

Consider the following problem, where $C \subseteq V$ is any clique :

$$\max_{\text{s.t } x \in \Delta^{(C)}} h(x) \tag{115}$$

Proposition 4.4.1. *The unique local (hence, global) minimizer of* (115) *is* $x^{(C)}$.

Proof. For $x \in \Delta^{(C)}$, $h(x) = L_{\overline{G}}(x) + \Phi(x) = \Phi_C(x_C)$ because $\text{supp}(x) = C$ is a clique. As $\Delta^{(C)}$ is convex and Φ_C is strictly convex, there exists a unique minimizer of h on $\Delta^{(C)}$, which is necessarily $x^{(C)}$. Indeed, denote x^* this minimizer. As Φ is permutation invariant, we must have $\mathcal{P}(x^*) \cap \Delta^{(C)} = \{x^*\}$ i.e. all the indices in C are equal, and because $\sum_i x_i = 1$, necessarily $x^* = x^{(C)}$. $\square$

The previous proposition shows that the unique maximizer in a face whose support is a clique is the characteristic vector of that clique. The next step is to prove that when we join these faces whose supports are cliques, the only maximizers remaining are the characteristic vectors of maximal cliques. Thus consider the following problem :

$$\min_{\text{s.t. } x \in \Delta^0} h(x) \tag{116}$$

Proposition 4.4.2. *A point $x \in \Delta^0$ is a local minimizer of* (116) *if and only if $x = x^{(C)}$ for some maximal clique C. Moreover, every local minimizer of* (116) *is strict.*

Proof. $\Rightarrow$: First, observe that, for any local minimizer x of (116), by definition of Δ^0, there exists some maximal clique C such that $x \in \Delta^{(C)}$ which implies that $x = x^{(C)}$ by the previous proposition.

$\Leftarrow$: Let C be a maximal clique, and suppose by way of contradiction that $x^{(C)}$ is not a strict minimizer in (116). Then, for every $k \in \mathbb{N}^*$, there exists some $x^k \in \Delta^0$ with $0 < \|x^k - x^{(C)}\|_2 < 1/k$ such that $h(x^k) \leq h(x^{(C)})$. Because there are only finitely many sets in the union of (112), there must exist some clique C' and some subsequence $(x^{k_l})_{l=1}^{+\infty} \subseteq (x^k)_{k=1}^{+\infty}$ such that $x^{k_l} \in \Delta^{(C')}$ for each $l \geq 1$, with $x^{k_l} \to x^{(C)}$. Hence, $x^{(C)} \in \overline{\Delta^{(C')}} = \Delta^{(C')}$, which implies $C = \text{supp}(x^{(C)}) \subseteq C'$. Because C is maximal, we must have $C' = C$ and thus $x^{k_l} \in \Delta^{(C')} = \Delta^{(C)}$ for each $l \geq 1$. Thus $x^{(C)}$ is not a strict local maximizer of (115), contradicting 4.4.1. $\square$

The next lemma is essential in order to prove the correspondence between global maxima of the continuous problem and the maximum cliques of the graph.

Lemma 4.4.3. *If C^1 and C^2 are cliques, then*

$$|C^1| < |C^2| \iff h(x^{(C^1)}) > h(x^{(C^2)}) \tag{117}$$

Proof. Let C^1, C^2 be cliques.

$\Rightarrow$: Assume that $|C^1| < |C^2|$. Let C be any clique such that $C \subset C^2$ and $|C| = |C^1|$. Then $x^{(C^1)} \in \mathcal{P}(x^{(C)})$. Therefore, $h(x^{(C^1)}) = h(x^C)$.

Moreover by 4.4.1, $h(x^{(C)}) > h(x^{(C^2)})$, because $x^{(C^2)} \in \Delta^{(C^2)}$. Hence, $h(x^{(C^1)}) > h(x^{(C^2)})$.

$\Leftarrow$: Assume that $h(x^{(C^1)}) > h(x^{(C^2)})$. By contraposition of the previous part, we must have $|C^1| \leq |C^2|$. Moreover, if $|C^1| = |C^2|$, then $x^{(C^1)} \in \mathcal{P}(x^{(C^2)})$ and thus $h(x^{(C^1)}) = h(x^{(C^2)})$, a contradiction. $\square$

The proof of the one-one correspondence between local / global maxima and characteristic vectors of maximal / maximum cliques on the restricted search space Δ^0 is now straightforward.

Proposition 4.4.4. *A point $x \in \Delta^0$ is a global minimizer of* (116) *if and only if $x = x^{(C)}$ for some maximum clique C.*

Proof. Let $x \in \Delta^0$. Then, x is a global minimizer if and only if x is a local minimizer and $h(x) \leq h(\tilde{x})$ for every local minimizer $\tilde{x} \neq x$, which by 4.4.2, holds if and only if $x = x^{(C)}$ for some maximal clique C and $h(x^{(C)}) \leq h(x^{\tilde{C}})$ for every maximal clique $\tilde{C} \neq C$. The corollary then follows from the previous proposition. $\square$

So far, we proved the one-one correspondence between local / global minimizers and characteristic vectors of maximal / maximum cliques solely on the union of the faces whose support are cliques. The last step is to show that there is no minimizer lying outside these faces, and this is where condition **(C3)** becomes necessary.

Proposition 4.4.5. *Every local minimizers of* (114) *are in* Δ^0.

Proof. Let x be a minimizer of (114). If $x \in \Delta^0$ we are done. Therefore assume instead that $x \notin \Delta^0$. Then $\mathrm{supp}(x)$ is not a clique, and there exists an edge $\tilde{e} \in \overline{E}$ such that for all $k \in \tilde{e}$, $k \in \mathrm{supp}(x)$. Let $i, j \in \tilde{e}$ such that $x_i \leq x_j \leq \min_{k \in \tilde{e} \setminus \{i,j\}} x_k$ and define $x(t) := x + t(\mathbf{e}_i - \mathbf{e}_j)$ for $-x_i \leq t \leq x_j$.

Because $d = t(\mathbf{e}_i - \mathbf{e}_j)$ is feasible for $-x_i \leq t \leq x_j$, we have that $\nabla h(x)^T d = 0$, and taking a Taylor's expansion around x we get

$$h(x(t)) - h(x) = \frac{1}{2} d^T \nabla^2 h(x) d + \mathcal{O}(t^3) \tag{118}$$

$$= \frac{t^2}{2} \left[-2 \sum_{e \in \overline{E}} 1_{i,j \in e} \prod_{l \in e \setminus \{i,j\}} x_l + (\mathbf{e}_i - \mathbf{e}_j)^T \nabla^2 \Phi(x) (\mathbf{e}_i - \mathbf{e}_j) \right] + \mathcal{O}(t^3) \tag{119}$$

First,

$$-2 \sum_{e \in \overline{E}} 1_{i,j \in e} \prod_{l \in e \setminus \{i,j\}} x_l \leq -x_i^{k-2} - x_j^{k-2} \tag{120}$$

because at least edge $\tilde{e}$ contributes to one non-null term in the sum and all of the x_l are greater than x_i and x_j, and then by assumption **(C3)**

$$(\mathbf{e}_i - \mathbf{e}_j)^T \nabla^2 \Phi(x) (\mathbf{e}_i - \mathbf{e}_j) < x_i^{k-2} + x_j^{k-2} \tag{121}$$

Thus $h(x(t)) - h(x) < 0$ when t is sufficiently small i.e. there is no neighborhood of x in which $h(x)$ is a minimum thus x is not a minimizer, a contradiction. $\square$

We have that local and global minimizers are characteristic vectors of respectively maximal and maximum cliques, we now have to prove that all of these characteristic vectors are themselves minimizers.

Proposition 4.4.6. *Let C be a maximal clique, then $x^{(C)}$ is a local minimizer of* (114).

Proof. $x^{(C)}$ is a local minimizer if and only if strict complementarity stands, i.e. $\lambda_i(x^{(C)}) = 0$ for $i \in \mathrm{supp}(x^{(C)})$ and $\lambda_i(x^{(C)}) > 0$ for $i \notin \mathrm{supp}(x^{(C)})$.

Case 1 : $C = V$ i.e. the clique is the whole graph. As there is no $i \notin C$, it is sufficient to check that $\lambda_i(x^{(C)}) = 0$ for all i. Let $i \in C = \mathrm{supp}(x^{(C)})$, we have

$$\lambda_i(x^{(C)}) = \left(\nabla h(x^{(C)})\right)_i - \nabla h(x^{(C)})^T x^{(C)} = \left(\nabla h(x^{(C)})\right)_i - \sum_{l \in C} \left(\nabla h(x^{(C)})\right)_l x_l^{(C)} \tag{122}$$

$$= \left(\nabla h(x^{(C)})\right)_i - \left(\nabla h(x^{(C)})\right)_i \sum_{l \in C} x_l^{(C)} \tag{123}$$

$$= 0 \tag{124}$$

where we used the symmetry of Φ and $x_i^{(C)} = x_l^{(C)}$ for all l in (123) and $\sum_l x_l = 1$ for all $x \in \Delta$ in (124), thus we do have strict complementarity in this case.

Case 2 : $C \neq V$ i.e. there is at least one missing node. In this case, strict complementarity is equivalent to

$$\nabla h(x^{(C)})^T(\mathbf{e}_i - \mathbf{e}_j) > 0 \text{ for } i \notin C \text{ and } j \in C \tag{125}$$

Indeed, $\lambda_i(x^{(C)}) = \nabla h(x^{(C)})^T(\mathbf{e}_i - x^{(C)})$ so $\nabla h(x^{(C)})^T(\mathbf{e}_i - \mathbf{e}_j) = \lambda_i(x^{(C)}) - \lambda_j(x^{(C)})$ and for $j \in C = \mathrm{supp}(x^{(C)})$, we already have $\lambda_j(x^{(C)}) = 0$ by the same proof as for **Case 1**.

Now let $i \notin C$ and $j \in C$ and define $\tilde{x} := x^{(C)} + (e_i - e_j)/(2|C|)$. Then

$$\nabla L_{\overline{G}}(\tilde{x})^T(e_i - e_j) = \sum_{e \in \overline{E}} \mathbf{1}_{i \in e} \prod_{l \in e \setminus \{i\}} \tilde{x}_l - \sum_{e \in \overline{E}} \mathbf{1}_{j \in e} \prod_{l \in e \setminus \{j\}} \tilde{x}_l > 0 \tag{126}$$

because $i \notin C$ and there is at least one edge in $\overline{E}$ between $k-1$ vertices of C and i (or else the clique would not be maximal), and $j \in C$ so the second sum is null. Then by condition **(C2)** we get

$$\nabla \Phi(\tilde{x})^T(e_i - e_j) = \frac{\partial h}{\partial x_i}(\tilde{x}) - \frac{\partial h}{\partial x_j}(\tilde{x}) = \frac{\partial h}{\partial x_i}(\tilde{x}) - \frac{\partial h}{\partial x_i}(\tilde{x}) = 0 \tag{127}$$

because $\tilde{x}_i = \tilde{x}_j$.

Now define $\overline{h}(t) := h(x(t))$ where $x(t) := x^{(C)} + t(\mathbf{e}_i - \mathbf{e}_j)$ for $0 \leq t \leq x_j = 1/|C|$, so that $\overline{h}'(1/2|C|) = \left(\nabla L_{\overline{G}}(\tilde{x}) + \Phi(\tilde{x})\right)^T (\mathbf{e}_i - \mathbf{e}_j) > 0$. For $0 \leq t \leq x_j = 1/|C|$, we have

$$\begin{aligned}
\overline{h}''(t) &= (\mathbf{e}_i - \mathbf{e}_j)^T \nabla^2 h(x(t))(\mathbf{e}_i - \mathbf{e}_j) && (128)\\
&= -2 \sum_{e \in \overline{E}} 1_{i,j \in e} \prod_{l \in e \setminus \{i,j\}} x(t)_l + (\mathbf{e}_i - \mathbf{e}_j)^T \nabla^2 \Phi(x(t))(\mathbf{e}_i - \mathbf{e}_j) && (129)\\
&< 0 && (130)
\end{aligned}$$

by the same reasoning as in the proof of 4.4.5. Thus $\overline{h}'$ is strictly decreasing, so $\overline{h}'(0) > \overline{h}'(1/2|C|) > 0$. This concludes the proof since $\nabla h(x^{(C)})^T(\mathbf{e}_i - \mathbf{e}_j) = \overline{h}'(0)$. ☐

The one-one correspondence ”local / global maximizer $\longleftrightarrow$ characteristic vector of maximal / maximum cliques” is now established.

4.5 Particular cases

Mr. Zeffiro in his notes then suggested I study the following particular case, in order to simplify the conditions **(C1)** - **(C3)** under additional hypotheses.

4.5.1 Sum of real functions

Assume $k > 2$ and that $\Phi(x) = \sum_{i=1}^n q_i(x_i)$ for some $q_i \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$ i.e. that the regularizer is a sum of twice derivable real functions depending on only one dimension of x.

Lemma 4.5.1. *All the q_i are equal up to a constant, i.e. there exists $q : \mathbb{R} \to \mathbb{R}$ such that for all $1 \leq i \leq n$, $q_i' = q'$.*

Proof. Follows from condition **(C2)**. ☐

This first lemma shows that necessarily, Φ can be rewritten as $\Phi(x) = \sum_{i=1}^n q(x_i) + C$ where C is a constant.

Lemma 4.5.2. *Necessarily,* $q''(0) = 0$.

Proof. By condition **(C1)**, $q \in \mathcal{C}^2(\mathbb{R})$ and $q''(y) > 0$ when $y > 0$. Taking the limit $x \to 0$, we get $q''(0) \geq 0$. Now for $0 < t \leq \frac{1}{2}$ define $x(t) := (t, t, 1-2t, 0, \ldots, 0) \in \Delta$. By condition **(C3)**, we have $2q''(t) \leq 2t^{k-2}$. Taking the limit $t \to 0$ gives us $q''(0) \leq 0$ thus $q''(0) = 0$. □

The next proposition simplifies condition **(C3)**, allowing to study the regularizer on only one dimension instead of two at a time.

Proposition 4.5.3. *In this case, conditions **(C1)** - **(C3)** can be rewritten as*

> ***(C'1)*** $q \in \mathcal{C}^2(\mathbb{R})$ *and* $q''(y) > 0$ *when* $y > 0$
>
> ***(C'2)*** $q''(y) < y^{k-2}$ *for* $0 < y \leq \frac{1}{2}$ *and* $q''(y) \leq y^{k-2}$ *for* $\frac{1}{2} < y \leq 1$

Proof. $\Leftarrow$: if q satisfies **(C'1)** and **(C'2)**, then $\Phi(x) = \sum_{i=1}^{n} q(x_i)$ obviously satisfies **(C1)** - **(C3)**.

$\Rightarrow$: Assume Φ satisfies **(C1)** - **(C3)**. Then for all $x \in \Delta$, $\frac{\partial^2 \Phi}{\partial x_i^2}(x) = q''(x_i)$ so q satisfies **(C'1)**.

Let $y \in (0,1)$ and for $0 \leq t < 1-y$ define $x(t) := (y, 1-y-t, t, 0, \ldots, 0) \in \Delta$. By **(C3)**, we have that for all t, $q''(y) + q''(1-y-t) < y^{k-2} + (1-y-t)^{k-2}$. Taking the limit $t \to 1-y$, we get $q''(y) \leq y^{k-2}$. This inequality can be extended to $y = 1$ by continuity of q''. Now assume by contradiction there exists $y \in (0, \frac{1}{2}]$ such that $q''(y) = y^{k-2}$. Then applying **(C3)** with $x = (y, y, 1-2y, 0, \ldots, 0) \in \Delta$, we get $2y^{k-2} = 2q''(y) < 2y^{k-2}$, a contradiction. □

4.5.2 Generalization of the regularizers found in [15]

This subsection is dedicated to the adaptation of the three original regularizers given in [15], that were :

- $\phi(x) = \frac{1}{2}\|x\|_2^2$
- $\alpha_1\|x + \epsilon 1\|_p^p$ with $\epsilon \geq 0$, $p \geq 2$ and $0 \leq \alpha_1 \leq \frac{2}{p(p-1)(1+\epsilon)^{p-2}}$
- $\phi(x) = \alpha_2 \sum_{i=1}^{n}(e^{-\beta x_i} - 1)$, with $\beta > 0$ and $0 \leq \alpha_2 \leq \frac{2}{\beta^2}$ (this is an approximation of $-\alpha_2\|x\|_0$)

We claim that the two following regularizers satisfy conditions **(C1)** - **(C3)** :

- $\Phi_B(x) := \alpha\|x\|_k^k$ with $0 < \alpha < \frac{1}{k(k-1)}$
- $\Phi_1(x) := \alpha_1\|x + \epsilon\mathbf{1}\|_p^p - \frac{n}{2}\alpha_1 p(p-1)\epsilon^{p-2}x^2$ with $\epsilon > 0$, $p \geq k$ and $0 < \alpha_1 \leq \frac{1}{p(p-1)(1+\epsilon)^{p-2}}$

Proof. These two regularizers are sum of twice derivable functions depending on only one dimension of x, the results of the previous part can be applied.

- Φ_B : Obvious.
- Φ_1 : For all $x \in \Delta$, $\Phi_1(x) = \sum_{i=1}^{n}\left(\alpha_1(x_i+\epsilon)^p - \frac{1}{2}\alpha_1 p(p-1)\epsilon^{p-2}x^2\right)$. For $y \in [0,1]$, denote $q(y) = \alpha_1(y+\epsilon)^p - \frac{1}{2}\alpha_1 p(p-1)\epsilon^{p-2}x^2$. $q''(y) = \alpha_1 p(p-1)(y+\epsilon)^{p-2} - \alpha_1 p(p-1)\epsilon^{p-2} > 0$ when $y > 0$ and for $y > 0$, $q''(y) < \alpha_1 p(p-1)(y^{p-2} + \epsilon^{p-2}) - \alpha_1 p(p-1)\epsilon^{p-2} = \alpha_1 p(p-1)y^{p-2} \leq y^{k-2}$.

□

4.5.3 Future work

The regularizer $\Phi_2(x) := \alpha_2 \sum_{i=1}^n \left(e^{-\beta x_i} - 1\right)$ with $\beta > 0$, and $0 < \alpha_2 < \frac{1}{\beta^2}$ of [15] was difficult to adapt to this new setting, so I wanted (but did not before the end of my internship) to replace it by an approximation on $[0,1]$ and under the norm $\|\cdot\|_2$ of the 0-norm (which was already approximated by $y \to \alpha_2(e^{-\beta y} - 1)$), with the form $y \to ay^k + by + c$ under the constraint $0 < a \leq \frac{1}{k(k-1)}$ and $\Phi_2(0) = -1$. The problem can be reformulated as

$$\begin{aligned} &\min_{a,b,c} \int_0^1 \left(ay^k + by + c\right)^2 dy \\ &\text{s.t. } 0 < a \leq \frac{1}{k(k-1)}, \quad c = -1 \end{aligned} \tag{131}$$

$$\iff \tag{132}$$

$$\begin{aligned} &\min_{a,b,c} \left[\frac{a^2}{2k+1} + \frac{b^2}{2} + c^2 + \frac{a^2}{2k+1} + 2\frac{ab}{k+2} + 2\frac{ac}{k+1} + 2\frac{bc}{2}\right] \\ &\text{s.t. } 0 < a \leq \frac{1}{k(k-1)}, \quad c = -1 \end{aligned} \tag{133}$$

Then the implementation in the first part (MsdCP) can easily be adapted to this problem. As in [15], extensive comparisons of the performance of the regularizers on the graphs of the DIMACS dataset need to be done.

5 Discrete / continuous problems equivalence via Lovász extensions

5.1 Introduction

The process of transforming combinatorial problems into continuous ones has been studied using mathematical constructs known as Lovász extensions. These extensions map a set to its characteristic vector and then continuously extend the underlying function across the entire space. I first read [20], [17], [18], and [19], then wrote a summary of these papers, which can be accessed here.

The objective was to get an understanding of how these techniques allow to retrieve continuous formulations for problems like the MCP on graphs and hypergraphs, in order to apply them to the MsdCP, which, at present, needs a two-step optimization process.

This section begins with definitions and follows with key results, a final part is dedicated to various applications, starting with the classical MCP in graphs (for $k = 2$). While only the last results were used, I left unused parts to showcase the progression of ideas behind these extensions, in order to help understanding.

5.2 Definitions and notations

Let $V := \{1, \dots, n\}$ be a finite and non-empty set, and denote $\mathcal{P}(V)$ its power set. Define also $\mathcal{P}(V)^k := \{(S_1, \dots, S_k) : S_i \subset V\}$ and $\mathcal{P}_k(V) := \{(S_1, \dots, S_k)\} \in \mathcal{P}(V)^k : S_i \cap S_j = \emptyset\}$.

We say that two vectors $x, y \in \mathbb{R}^n$ are comonotonic if $(x_i - x_j)(y_i - y_j) \geq 0$ for all i, j i.e. they have the same "variations".

A discrete function $F : \mathcal{A} \to \mathbb{R}$ is submodular if $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$ for all $A, B \in \mathcal{A}$. A continuous function $F : \mathbb{R}^n \to \mathbb{R}$ is submodular if $F(x) + F(y) \geq F(\min(x, y)) + F(\max(x, y))$ where the min and max are taken elementwise.

5.3 Lovász extensions

5.3.1 Definitions

Given a function $f : \mathcal{P}(V) \to \mathbb{R}$, its Lovász extension extends the domain of f to $\mathbb{R}^n$:

> Denote $x = (x_1, \dots, x_n) \in \mathbb{R}$, and let $\sigma : V \cup \{0\} \to V \cup \{0\}$ be a permutation such that $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$ and $\sigma(0) := 0$ where we add $x_0 := 0$ to x. The Lovász extension is defined as
>
> $$f^L(x) := \sum_{i=0}^{n-1} (x_{\sigma(i+1)} - x_{\sigma(i)}) f(V^{\sigma(i)}(x)) \tag{134}$$
>
> where $V^0 := V$ and $V^{\sigma(i)}(x) := \{j \in V : x_j > x_{\sigma(i)}\}$.

The above formula can be rewritten as an integral :

$$f^L(x) = \int_{\min x_i}^{\max x_i} f(V^t(x))dt + f(V) \min x_i \tag{135}$$

where $V^t(x) := \{i \in V, x_i > t\}$ and with the additional assumption $f(\emptyset) = 0$,

$$f^L(x) = \int_{-\infty}^{0} (f(V^t(x)) - f(V))dt + \int_{0}^{+\infty} f(V^t(x))dt \tag{136}$$

5.3.2 Key properties

Proposition 5.3.1. *For a set S, denoting its characteristic vector $\mathbf{1}_S \in \{0,1\}^n$, $f^L(\mathbf{1}_S) = f(S)$.*

Proposition 5.3.2. *f^L is the unique function that is affine on each polyhedral cone $\mathbb{R}^n_\sigma := \{x \in \mathbb{R}^n, x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}\}$ and for which $f^L(\mathbf{1}_S) = f(S)$ for every set S.*

Proposition 5.3.3. *f^L is positively one-homogeneous, piecewise-linear and Lipschitzian continuous.*

Proposition 5.3.4. *f is submodular $\iff$ f^L is convex $\iff$ f^L is submodular.*

5.4 k-way extensions

5.4.1 Definitions

Lovász extensions transorm discrete functions taking one set as argument, but this can be generalized to functions taking multiple sets as arguments through k-way Lovász extensions. For a function $f : \mathcal{P}(V_1) \times \cdots \times \mathcal{P}(V_k) \to \mathbb{R}$, it is defined as

$$\begin{aligned} f^L(x^1, \dots, x^k) :&= \int_{\min x}^{\max x} f(V_1^t(x^1), \dots, V_k^t(x^k))dt + f(V_1, \dots, V_k) \min x & (137) \\ &= \int_{-\infty}^{0} (f(V_1^t(x^1), \dots, V_k^t(x^k)) - f(V_1, \dots, V_k))dt + \int_{0}^{+\infty} f(V_1^t(x^1), \dots, V_k^t(x^k))dt & (138) \end{aligned}$$

where the V_i^t are defined in the same way as for the classical case. Beware that the x^k are all vectors of $\mathbb{R}^n$. We define $\mathcal{D}_\mathcal{A} := \{x \in \mathbb{R}_+^{kn} | ((V^t(x^1), \dots, V^t(x^k)) \in \mathcal{A} \text{ for all } t \in \mathbb{R}\}$.

k-way submodularity is defined as follows :

> Given a tuple $V = (V_1, \dots, V_k)$ of finite sets and $\mathcal{A} \subset \{(A_1, \dots, A_k) | A_i \subset V_i\}$, a discrete function $f : \mathcal{A} \to \mathbb{R}$ is k-way submodular if
>
> $$\begin{aligned} &f(A_1, \dots, A_k) + f(B_1, \dots, B_k) \geq \\ &f(\min(A_1, B_1), \dots, \min(A_k, B_k)) + f(\max(A_1, B_1), \dots, \max(A_k, B_k)) \end{aligned} \tag{139}$$

5.4.2 Key properties

Proposition 5.4.1. *For the k-way Lovász extension f^L, we have*

(a) $f^L(\cdot)$ is positively one-homogeneous, piecewise linear and Lipschitz continuous

(b) $(\lambda f)^L = \lambda f^L$

Proposition 5.4.2. *For k-way Lovász extension, if f has a separable summation form $f(S_1, \ldots, S_k) = \sum_{i=0}^{k} f_i(S_i)$, we have $f^L(x^1, \ldots, x^k) = \sum_{i=1}^{k} f_i^L(x^i)$ for all $(x^1, \ldots, x^k)$.*

Proposition 5.4.3. *The following statements are equivalent :*

a) f is k-way submodular on $\mathcal{A}$

b) the k-way Lovász extension f^L is convex on each convex subset of $\mathcal{D}_{\mathcal{A}}$

c) the k-way Lovász extension f^L is submodular on $\mathcal{D}_{\mathcal{A}}$

5.5 Combinatorial and continuous optimization links via Lovász extensions

5.5.1 Definitions

Here, f^L denote any of the different Lovász extensions defined above. $\mathcal{A}$ is a restricted family of $\mathcal{P}(V)^k$.

5.5.2 Key results

Theorem 5.5.1. *Given set functions $f_1, \ldots, f_n : \mathcal{A} \to \mathbb{R}^+$ and a zero-homogeneous function $H : (\mathbb{R}^+)^m \setminus \{0\} \to \mathbb{R} \cup \{+\infty\}$ with $H(a+b) \geq \min(H(a), H(b))$ for all $a, b \in (\mathbb{R}^+)^m \setminus \{0\}$, we have*

$$\min_{S \in \mathcal{A}'} H(f_1(S), \ldots, f_n(S)) = \inf_{x \in \mathcal{D}'} H(f_1^L(x), \ldots, f_n^L(x)) \tag{140}$$

where $\mathcal{A}' = \{S \in \mathcal{A} : (f_1(S), \ldots, f_n(S)) \in Dom(H)\}$, $\mathcal{D}' = \{x \in \mathcal{D}_{\mathcal{A}} \cap (\mathbb{R}^+)^V : (f_1^L(x), \ldots, f_n^L(x)) \in Dom(H)\}$ and $Dom(H) = \{a \in (\mathbb{R}^+)^m \setminus \{0\} : H(a) \in \mathbb{R}\}$.

Proposition 5.5.2. *Given two set functions $f, g : \mathcal{A} \to [0, \infty)$, let $\tilde{f}, \tilde{g} : \mathcal{D}_{\mathcal{A}} \to \mathbb{R}$ satisfying $\tilde{f} \geq f^L$, $\tilde{g} \leq g^L$, $\tilde{f}(\mathbf{1}_S) = f(S)$ and $\tilde{g}(\mathbf{1}_S) = g(S)$. Then*

$$\min_{S \in \mathcal{A} \cap supp(g)} \frac{f(S)}{g(S)} = \inf_{\Psi \in \mathcal{D}_{\mathcal{A}} \cap supp(\tilde{g})} \frac{\tilde{f}(\Psi)}{\tilde{g}(\Psi)} \tag{141}$$

If we replace the conditions by $\tilde{f} \leq f^L$, $\tilde{g} \geq g^L$. Then

$$\max_{S \in \mathcal{A} \cap supp(g)} \frac{f(S)}{g(S)} = \sup_{\Psi \in \mathcal{D}_{\mathcal{A}} \cap supp(\tilde{g})} \frac{\tilde{f}(\Psi)}{\tilde{g}(\Psi)} \tag{142}$$

Proposition 5.5.3. *Let $f, g : \mathcal{A} \to [0, +\infty)$ be two set functions and $f = f_1 - f_2$ and $g = g_1 - g_2$ be decompositions of differences of submodular functions. Let $\tilde{f}_2, \tilde{g}_1$ be the restrictions of positively one-homogeneous convex functions onto $\mathcal{D}_{\mathcal{A}}$, with $f_2(S) = \tilde{f}_2(\mathbf{1}_S)$ and $g_1(S) = \tilde{g}_1(\mathbf{1}_S)$. Define $\tilde{f} := f_1^L - \tilde{f}_2$ and $\tilde{g} := \tilde{g}_1 - g_2^L$. Then,*

$$\min_{S \in \mathcal{A} \cap supp(g)} \frac{f(S)}{g(S)} = \min_{x \in \mathcal{D}_{\mathcal{A}} \cap supp(\tilde{g})} \frac{\tilde{f}(S)}{\tilde{g}(S)} \tag{143}$$

5.6 Homogeneous and piecewise multilinear extensions

5.6.1 Definitions

For now, Lovász extensions satisfy the property $(f+g)^L = f^L + g^L$, and homogeneous and piecewise multilinear extensions extend this property to $(gf)^L = g^L f^L$, sacrificing the equalities between the discrete and continuous problems above and replacing them by inequalities, the reverse inequalities having to be proved on a case by case basis.

Given $V_i := \{1, \ldots, n_i\}$ and the power set $\mathcal{P}(V_i)$ for $i = 1, \ldots, k$, for a discrete function $f : \mathcal{P}(V_1) \times \cdots \times \mathcal{P}(V_k) \to \mathbb{R}$, its piecewise multilinear extension is defined on $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ by

$$f^M(x^1, \ldots, x^k) := \sum_{i_1 \in V_1, \ldots, i_k \in V_k} \prod_{l=1}^{k} (x^l_{\sigma_l(i_l)} - x^l_{\sigma_l(i_{l-1})}) f(V^{i_1}(x^1), \ldots, V^{i_k}(x^k)) \tag{144}$$

where $V^i(x^l) := \{j \in V_l | x^l_j > x^l_{l-1}\}$ for $i \geq 2$, $V^1(x^l) := V_l$, σ_l is a permutation of indices sorting x_l by non-decreasing order, and we add $x^l_0 := 0$ to each x^l.

For such a piecewise multilinear extension, its piecewise polynomial extension is defined for all $x \in \mathbb{R}$ as

$$f^M_\Delta(x) := f^M(x, \ldots, x) \tag{145}$$

Under the assumption that $f(A_1, \ldots, A_k) = 0$ whenever $A_i \in \{V, \emptyset\}$ for some i, we have the following integral representation

$$f^M(x^1, \ldots, x^k) = \int_{\min x^k}^{\max x^k} \cdots \int_{\min x^1}^{\max x^1} f(V^{t_1}(x^1), \ldots, V^{t_k}(x^k)) dt_1 \ldots dt_k \tag{146}$$

where $V^{t_l}(x^l) := \{j \in V | x^l_j > t_l\}$.

5.6.2 Key properties

Proposition 5.6.1. *Let $f : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}$ and define $\tilde{f} : A \in \mathcal{P}(V_1) \to f^M(\mathbf{1}_A, x^2, \ldots, x^k) \in \mathbb{R}$. Then $\tilde{f}^L(x) = f^M(x, x^2, \ldots, x^k)$ for any $x \in \mathbb{R}^{n_1}$.*

Proposition 5.6.2. *A function $f : \mathcal{P}(V_1) \times \ldots \mathcal{P}(V_k) \to \mathbb{R}$ is modular on each component if and only if f^M is multilinear. Then we have*

$$f^M(x^1, \ldots, x^k) = \int_0^{\max x^k} \cdots \int_0^{\max x^1} f(V^{t_1}(x^1), \ldots, V^{t_k}(x^k)) dt_1 \ldots dt_k \tag{147}$$

where $V^{t_l}(x^l) = \{j \in V | x^l_j > t_l\}$.

Proposition 5.6.3. *If f is of the form $f(A_1, \ldots, A_k) = \prod_{i=1}^k f_i(A_i)$ then $f^M(x^1, \ldots, x^k) = \prod_{i=1}^k f_i^L(x^i)$.*

5.7 Combinatorial and continuous optimization links via piecewise multilinear extensions

5.7.1 Definitions

Given constraints sets $\mathcal{A} \subset (\mathcal{P}(V) \backslash \{\emptyset\})^k$ (or $\mathcal{A} \subset (\mathcal{P}_2(V) \backslash \{\emptyset, \emptyset\})^k$) and $\mathcal{D} \subset (\mathbb{R}^n_+)^k$, the feasible sets $\mathcal{A}(\mathcal{D})$ and $\mathcal{D}(\mathcal{A})$ are defined in the following way : $\mathcal{D}(\mathcal{A}) := \{(x^1, \ldots, x^k) \in (\mathbb{R}^n_+)^k | (V^{t_1}(x^1), \ldots, V^{t_k}(x^k)) \in \mathcal{A}$, for all $t_i < \max x^i\}$ and $\mathcal{A}(\mathcal{D}) := \{(V^{t_1}(x^1), \ldots, V^{t_k}(x^k)) \in (\mathcal{P}(V) \setminus \{\emptyset\})^k$, for all $(x^1, \ldots, x^k) \in (\mathbb{R}^n_+)^k, t_1, \ldots, t_k \in \mathbb{R}\}$. $(\mathcal{A}, \mathcal{D})$ is called a perfect domain pair if $\mathcal{A} = \mathcal{A}(\mathcal{D})$ and $\mathcal{D} = \mathcal{D}(\mathcal{A})$.

5.7.2 Key results

Theorem 5.7.1. *Given $f : \mathcal{A} \to \mathbb{R}$ and $g : \mathcal{A} \to [0, +\infty)$, we have*

$$\sup_{A \in \mathcal{A} \cap supp(g)} \frac{f(A)}{g(A)} \leq \sup_{x \in \mathcal{D} \cap supp(g^M)} \frac{f^M(x)}{g^M(x)} \leq \sup_{A \in \tilde{\mathcal{A}}} \frac{f(A)}{g(A)} \tag{148}$$

whenever $\{\mathbf{1}_A : A \in \mathcal{A}\} \subset \mathcal{D}$ and $\mathcal{A}(D) \subset \tilde{\mathcal{A}}$. The above inequality still holds replacing sup *and $\leq$ by* inf *and $\geq$. If we further assume that $(\mathcal{A}, \mathcal{D})$ is a perfect domain pair, and $supp(f) \subset supp(g)$, then*

$$\max_{A \in \mathcal{A} \cap supp(g)} \frac{f(A)}{g(A)} = \max_{x \in \mathcal{D} \cap supp(g^M)} \frac{f^M(x)}{g^M(x)} \tag{149}$$

and the same holds replacing max *with* min.

Theorem 5.7.2. *Let $H : \mathbb{R}_+^* \to \mathbb{R} \cup \{+\infty\}$ be a zero-homogeneous and quasi-concave function. For any function $f_1, \ldots, f_n : \mathcal{A} \to \mathbb{R}_+$, we have*

$$\min_{A \in \mathcal{A}} H(f_1(A), \ldots, f_n(A)) = \inf_{x \in \mathcal{D}} H(f_1^M(x), \ldots, f_n^M(x)) \tag{150}$$

where $(\mathcal{A}, \mathcal{D})$ forms a perfect domain pair w.r.t. the piecewise multilinear extension.

In addition, if $H : \mathbb{R}_+^* \to \mathbb{R} \cup \{-\infty\}$ is a zero-homogeneous and quasi-convex function, for any function $f_1, \ldots, f_n : \mathcal{A} \to \mathbb{R}_+$, we have

$$\max_{A \in \mathcal{A}} H(f_1(A), \ldots, f_n(A)) = \sup_{x \in \mathcal{D}} H(f_1^M(x), \ldots, f_n^M(x)) \tag{151}$$

where $(\mathcal{A}, \mathcal{D})$ forms a perfect domain pair w.r.t. the piecewise multilinear extension.

Corollary 5.7.3. *For a log-concave polynomial P of degree d in n variables, and for $f_1, \ldots, f_, : \mathcal{A} \to [0, +\infty)$, we have*

$$\min_{A \in \mathcal{A}} \frac{P(f_1(A), \ldots, f_n(A)}{(f_1(A) + \cdots + f_n(A))^d} = \inf_{x \in \mathcal{D}} \frac{P(f_1^M(A), \ldots, f_n^M(A)}{(f_1^M(A) + \cdots + f_n^M(A))^d} \tag{152}$$

where $(\mathcal{A}, \mathcal{D})$.

5.8 Applications

5.8.1 Lemma

This first lemma (present in [18]) is essential to prove the equality between the discrete version of the MCP and its continuous unregularized one.

Lemma 5.8.1. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be smooth functions such that g is positive on $(\mathbb{R}_+^*)^n \setminus \{0\}$. For a maximizer (resp. minimizer) x of $\frac{f}{g}_{|(\mathbb{R}_+^*)^n \setminus \{0\}}$ (if it exists), let v be such that $x + v \in (\mathbb{R}_+^*)^n \setminus \{0\}$, $supp(v) \subset supp(x)$, $t \in \mathbb{R} \to g(x + tv)$ is constant, and $\frac{\partial^2 f}{\partial y_i \partial y_j}(y) = 0$ for all $i, j \in supp(v)$, for all $y \in \mathbb{R}^n$. If we further assume that f is real analytic, then $x + v$ is also a maximizer of $\frac{f}{g}_{|(\mathbb{R}_+^*)^n \setminus \{0\}}$.*

Proof. Let x be a critical point of $\frac{f}{g}_{|(\mathbb{R}_+^*)^n \setminus \{0\}}$ and let $v \in \mathbb{R}^n$ be such that $\langle \nabla g(x), v \rangle = 0$ and $\text{supp}(v) \subset \text{supp}(x)$, then $\langle \nabla f(x, v \rangle = 0$. Indeed, for any $i \in \text{supp}(x)$, we have $\frac{\partial}{\partial x_i} \frac{f}{g}(x) = 0$ because x is a critical point, thus $\frac{\partial f}{\partial x_i}(x) = \frac{f}{g}(x) \frac{\partial g}{\partial x_i}(x)$ for any $i \in \text{supp}(x)$. Because $v_i = 0$ whenever $i \notin \text{supp}(x)$, we have

$$\langle \nabla f(x), v \rangle = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}(x) = \sum_{i \in \text{supp}(x)} v_i \frac{\partial f}{\partial x_i}(x) \tag{153}$$

$$= \sum_{i \in \text{supp}(x)} v_i \frac{f}{g}(x) \frac{\partial g}{\partial x_i}(x) = \frac{f}{g}(x) \langle \nabla g(x), v \rangle = 0 \tag{154}$$

Now we prove the lemma. It follows from $g(x+tv) = g(x)$ for all $t \in \mathbb{R}$ that $\langle \nabla g(x), v\rangle = 0$, and thus by the above claim, we have $\langle \nabla f(x), v\rangle = 0$. Since f is a real analytic function, $t \to f(x+tv)$ must be real analytic. Note that $\frac{df(x+tv)}{dt}(t=0) = \langle \nabla f(x), v\rangle = 0$ and for any $k \geq 2$,

$$\frac{d^k f(x+tv)}{dt^k}(t=0) = \sum_{i_1,\dots,i_k=1}^{n} v_{i_1}\dots v_{i_k} \frac{\partial^k f}{\partial x_{i_1}\dots \partial x_{i_k}}(x) \tag{155}$$

$$= \sum_{i_1,\dots,i_k \in \mathrm{supp}(v)} v_{i_1}\dots v_{i_k} \frac{\partial^k f}{\partial x_{i_1}\dots \partial x_{i_k}}(x) = 0 \tag{156}$$

where the last equality is due to the condition that $\frac{\partial f}{\partial x_i \partial x_j}(x) = 0$ for all $i, j \in \mathrm{supp}(v)$.

Therefore, the real analytic function $t \to f(x+tv)$ is constant. This implies that $f(x+v) = f(x)$, and hence $\frac{f(x+v}{g(x+v)} = \frac{f(x)}{g(x)}$ thus $x+v$ is also a maximizer of $\frac{f}{g}_{|(\mathbb{R}_+^*)^n\setminus\{0\}}$.

The case of minimizer is similar. □

5.8.2 Application to the MCP on 2-graphs

The following proof is present in [18] and gives a second way to discover the continuous formulation and prove its equivalence. Here, $E(S,T)$ denote the number of edges $e := (i,j)$ of the graph where $i \in S$ and $j \in T$.

Proposition 5.8.2. *Let $G := (V, E)$ be an undirected graph with no edge between a node and itself. Denote A its adjacency matrix. The following problem*

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,S)|}{|S|^2} \tag{157}$$

has solutions that are maximum cliques. Indeed, for any set S,

$$\frac{|E(S,S)|}{|S|^2} \leq \frac{|S|(|S|-1)}{2|S|^2} \tag{158}$$

with equality only if S is a clique, and the above is maximized with S a maximum clique.

Define $f(S,T) := |E(S,T)|$ and $g(S,T) := |S||T|$. Then the piecewise multilinear extensions of f and g defined on $\mathbb{R}_+^n$ satisfy $f^M(x,y) = x^T Ay$ and $g^M(x,y) = \|x\|_1\|y\|_1$.

Then

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,S)|}{|S|^2} = \max_{x\neq 0} \frac{2\sum_{(i,j)\in E} x_i x_j}{\|x\|_1^2} \tag{159}$$

Proof. $g(S,T) = \tilde{g}(S)\tilde{g}(T)$ and $\tilde{g}$ is modular so $\tilde{g}^M(x) = \langle u, x\rangle^2$ where $u = (\tilde{g}(\{1\}),\dots,\tilde{g}(\{n\})) = \mathbf{1}$.

f is modular on each component so its extension f^M must be multilinear and thus $f^M(x,y) = x^T My$ where $M = (f(\{i\},\{j\}))_{n\times n} = A$.

A similar result to 5.7.1 holds for multilinear extensions evaluated diagonally, which gives us the inequality

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,S)|}{|S|^2} \leq \max_{x\in\mathbb{R}_+^n\setminus\{0\}} \frac{2\sum_{(i,j)\in E} x_i x_j}{\|x\|_1^2} = \max_{x\neq 0} \frac{2\sum_{(i,j)\in E} x_i x_j}{\|x\|_1^2} \tag{160}$$

For the equality, we have to show that there exists a set S such that $\mathbf{1}_S$ is a maximizer of the right side of (160). Let x be a maximizer of the right side in $\mathbb{R}_+^n \setminus \{0\}$. If $f(\{i\},\{j\}) = 0$ and $x_i x_j > 0$ for some $i \neq j$, then taking v defined as $v_i = -x_i$, $v_j = x_i$ and $v_l = 0$ for $l \neq i, j$, we have that $x+v$ is also a maximizer by 5.8.1, and repeating the process, we finally obtain a subset $S \subset V$ satisfying $\text{supp}(x) = S$ and $f(\{i\},\{j\}) > 0$ for all $i \neq j$ in S i.e. S is a clique. Then

$$\frac{f_\Delta^M(x)}{g_\Delta^M(x)} = \frac{x^T A x}{(\mathbf{1}^T x)^2} = 1 - \frac{\sum_{i\in S} x_i^2}{(\sum_{i\in S} x_i)^2} \leq 1 - \frac{1}{|S|} \tag{161}$$

and the equality holds if and only if $x_i = \text{Const}$ for $i \in S$. In consequence, $\mathbf{1}_S$ is a maximizer of the right side of (160). The proof is completed. $\square$

Remark 5.8.3. *There is no guarantee that the maximizer x^* of the right side of (164) has a meaning in terms of set (i.e. there is no guarantee that $x^* = \mathbf{1}_S$ for a particular set S). Such a solution is sometimes called a "spurious" solution, and here we only have that a set S maximizing the left side implies that $\mathbf{1}_S$ maximizes the right side.*

The authors of [9] added a regularizer $\alpha\|x\|_2^2$ with $\alpha \in (0,2)$ which in [15] was further relaxed to a function Φ satisfying $\nabla^2\Phi \geq 0$, $\nabla^2\phi < 2$ and Φ permutation invariant. The first inequality can be replaced by a strict one to ensure that not only the global maxima but also the local ones are in relation with the characteristic vectors of sets.

5.8.3 Application to the MCP on k-graphs induced by k-cliques of a 2-graph

The authors of [18] also gave a short proof for an extension of this result to k-graphs induced by k-cliques of a 2-graph.

Let $G^* := (V, E^*)$ be an undirected 2-graph, and let $G := (V, E)$ be the graph induced by its k-cliques, i.e. the graph mapping every k-cliques of G^* to an edge in G (for example for every triangle in a 2-graph, we map a 3-edge in the associated 3-graph). Denote $\mathcal{A}$ its adjacency tensor. Denote $E(S_1, \dots, S_k)$ the number of edges $e := (i_1, \dots, i_k)$ of the graph where $i_j \in S_j$ for all j.

Proposition 5.8.4. *The following problem*

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,\dots,S)|}{|S|^k} \tag{162}$$

has solutions that are maximum cliques. Indeed, for any set S,

$$\frac{|E(S,\dots,S)|}{|S|^k} \leq \frac{\binom{|S|}{k}}{|S|^k} \tag{163}$$

with equality only if S is a clique, and the above is maximized with S a maximum clique.

Define $f(S_1,\dots,S_k) := |E(S_1,\dots,S_k)|$ and $g(S_1,\dots,S_k) := |S_1|\dots|S_k|$. Then the piecewise multilinear extensions of f and g defined on $\mathbb{R}_+^n$ satisfy $f^M(x^1,\dots,x^k) = \mathcal{A}x^1\dots x^k$ and $g^M(x^1,\dots,x^k) = \|x^1\|_1\dots\|x^k\|_1$.

Then

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,\dots,S)|}{|S|^k} = \max_{x\neq 0} \frac{k!\sum_{e\in E}\prod_{i\in e} x_i}{\|x\|_1^k} \tag{164}$$

Proof. $g(S_1,\dots,S_k) = |S_1|\dots|S_k|$ and $\tilde{g}$ is modular so $\tilde{g}^M(x) = \langle u, x\rangle^k$ where $u = (\tilde{g}(\{1\}),\dots,\tilde{g}(\{n\})) = \mathbf{1}$.

f is modular on each component so its extension f^M must be multilinear and thus $f^M(x^1,\dots,x^k) = \mathcal{M}x^1\dots x^k$ where $\mathcal{M} = (f(\{i_1\},\dots,\{i_k\}))_{n\times\dots\times n} = \mathcal{A}$.

A similar result to 5.7.1 holds for multilinear extensions evaluated diagonally, which gives us the inequality

$$\max_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,\dots,S)|}{|S|^k} \leq \max_{x\in\mathbb{R}_+^n\setminus\{0\}} \frac{k!\sum_{e\in E}\prod_{i\in e} x_i}{\|x\|_1^k} = \max_{x\neq 0} \frac{k!\sum_{e\in E}\prod_{i\in e} x_i}{\|x\|_1^k} \tag{165}$$

For the equality, we have to show that there exists a set S such that $\mathbf{1}_S$ is a maximizer of the right side of (171). Let x be a maximizer of the right side in $\mathbb{R}^n_+ \setminus \{0\}$. If there exist $i, j \in \mathrm{supp}(x)$ such that for all edges $e = (e_1, \dots, e_k) \in E$ such that $i, j \in e$ and $x_{e_l} > 0$ for all l, we have $f(\{e_1\}, \dots, \{e_k\}) = 0$ (or in other words if we have two indices such that no k-cliques in the original graph contains them both), then taking v defined as $v_i = -x_i$, $v_j = x_i$ and $v_l = 0$ for $l \neq i, j$ where $i, j \in e$, we have that $x + v$ is also a maximizer by 5.8.1, and repeating the process, we finally obtain a subset $S \subset V$ satisfying $\mathrm{supp}(x) = S$ and $f(\{i_1\}, \dots \{i_k\}) > 0$ for all $i_1, \dots, i_k$ in S (remind that G is induced by k-cliques of G^*), i.e. S is a clique. Then by McLaurin inequality,

$$\frac{f^M_\Delta(x)}{g^M_\Delta(x)} = \frac{k! \sum_{e \in E} \prod_{i \in e} x_i}{(\mathbf{1}^T x)^k} = \frac{k! \sum_{e \in E(S)} \prod_{i \in e} x_i}{\|x\|_1^k} \tag{166}$$

$$\leq \frac{\binom{|S|}{k} \left(\frac{\|x\|_1}{|S|}\right)^k}{\|x\|_1^k} \tag{167}$$

and the equality holds if and only if $x_i = \mathrm{Const}$ for $i \in S$. In consequence, $\mathbf{1}_S$ is a maximizer of the right side of (171). The proof is completed. □

Remarks This result only holds for k-graphs induced by k-cliques of a 2-graphs. The impossibility to fully generalize this technique to k-graphs is discussed in the introduction of [10]. In the case of this proof, it is due to the inability to verify every hypothesis of the lemma at the beginning of the subsection.

5.8.4 Application to the MCP on k-graphs, based on the problem derived in [10]

Let $G := (V, E)$ be an undirected k-graph. Denote $\mathcal{A}(\overline{G})$ the adjacency tensor of the complement of G. Denote $\overline{E}(S_1, \dots, S_k)$ the number of edges $e := (i_1, \dots, i_k)$ of the complement of the graph where $i_j \in S_j$ for all j.

The following is an attempt to retrieve the known continuous formulation for the MCP for hypergraphs present in [10] through the techniques develop in the four papers. We managed to retrieve the formulation only partially.

Proposition 5.8.5. *The following problem*

$$\min_{S \in \mathcal{P}(V) \setminus \{\emptyset\}} \frac{|\overline{E}(S, \dots, S)| + \tau |S|}{|S|^k} \tag{168}$$

has solutions that are maximum cliques when $0 < \tau < \frac{1}{|V|(|V|^{k-1}-1)}$. Indeed, for any clique C and any set S such that S is not a clique,

$$\frac{0 + \tau |C|}{|C|^k} \leq \tau \leq \frac{1 + \tau |V|}{|V|^k} \leq \frac{|\overline{E}(S, \dots, S)| + \tau |S|}{|S|^k} \tag{169}$$

and the first term is minimized with C a maximal clique.

Define $f(S_1, \dots, S_k) := |E(S_1, \dots, S_k)| + \tau |S_1 \cap \dots \cap S_k|$ and $g(S_1, \dots, S_k) := |S_1| \dots |S_k|$. Then the piecewise multilinear extensions of f and g defined on $\mathbb{R}^n_+$ satisfy $f^M(x^1, \dots, x^k) = \mathcal{A}(\overline{G}) x^1 \dots x^k + \tau \|x\|_k^k$ and $g^M(x^1, \dots, x^k) = \|x^1\|_1 \dots \|x^k\|_1$.

Then

$$\min_{S \in \mathcal{P}(V) \setminus \{\emptyset\}} \frac{|\overline{E}(S, \dots, S)| + \tau |S|}{|S|^k} = \min_{x \neq 0} \frac{k! \sum_{e \in \overline{E}} \prod_{i \in e} x_i + \tau \|x\|_k^k}{\|x\|_1^k} \tag{170}$$

Proof. $g(S_1, \dots, S_k) = \tilde{g}(S_1) \dots \tilde{g}(S_k)$ and $\tilde{g}$ is modular so $g^M(x) = \langle u, x \rangle^k$ where $u = (\tilde{g}(\{1\}), \dots, \tilde{g}(\{n\})) = \mathbf{1}$.

f is modular on each component so its extension f^M must be multilinear and thus $f^M(x^1, \ldots, x^k) = \mathcal{M}x^1 \ldots x^k$ where $\mathcal{M} = (f(\{i_1\}, \ldots, \{i_k\}))_{n\times\cdots\times n} = \mathcal{A}(\overline{G}) + \tau\mathcal{I}$.

A similar result to 5.7.1 holds for multilinear extensions evaluated diagonally, which gives us the inequality

$$\min_{S\in\mathcal{P}(V)\setminus\{\emptyset\}} \frac{|E(S,\ldots,S)|}{|S|^k} \geq \min_{x\in\mathbb{R}^n_+\setminus\{0\}} \frac{k!\sum_{e\in E}\prod_{i\in e} x_i + \tau\|x\|_k^k}{\|x\|_1^k} \tag{171}$$

As $\left[0, \frac{1}{|V|(|V|^{k-1}-1)}\right] \subset \left[0, \frac{1}{k(k-1)}\right]$, the proof in [10] allows to get the reverse inequality. □

5.8.5 Application to the s-defective clique case

We also searched for a second continuous formulation for the s-defective clique case, which for now needs a two-steps optimization process. I suggested the following discrete function, but it is way too complicated to be used :

$$h(S) := (s+2)^{|S|}\left(s+1-\left(\frac{|S|(|S|-1)}{2} - |E(S,S)|\right)\right) \tag{172}$$

This can also be rewritten in fractional form :

$$h(S) = \frac{\left(s+1-\left(\frac{|S|(|S|-1)}{2} - |E(S,S)|\right)\right)}{\left(\frac{1}{s+2}\right)^{|S|}} \tag{173}$$

Remark 5.8.6. *The second factor can be rephrased as "$s+1$ - missing edges" and is there to ensure that there is indeed less than s missing edges (or else the factor would be equal to 0 or negative so not a solution).*

The first factor is to ensure that at a specific number of missing edges, the function is growing with the cardinal of the set. Its form is there to ensure that the tradeoff between the penalty of having additional missing edges and the increase in the first factor is always in favor of the latter, i.e. the largest decrease the second factor can have is $s+1 \to 1$ when going from a clique to a set with s missing edges i.e. dividing by $s+1$, so to ensure that an increase in the cardinal is profitable (even if we add s missing edges), we need the first factor to grow at least $s+2 > s+1$.

6 Conclusion

The work I did during my internship allowed to advance along three axes : we developed a novel continuous formulation for the MsdCP and studied the convergence of two algorithms able to solve it, as well as provided an implementation ; we extended the continuous formulation of the MCP in hypergraphs to a broader class of regularizers ; we laid the groundwork towards the use of continuous extensions to find continuous formulations for discrete problems of the same family as the MCP.

I will conclude by thanking Mr. Francesco Rinaldi and Mr. Damiano Zeffiro for their guidance during the six months I spent at the University of Padua.

References

[1] I. M. Bomze. "Evolution towards the maximum clique". In: *J. Global Optim.* 10.2 (1997), pp. 143–164.

[2] I. M. Bomze, M. Pelillo, and V. Stix. "Approximating the maximum weight clique using replicator dynamics". In: *IEEE Trans. Neural Network* 11.6 (2000), pp. 1228–1241.

[3] I. M. Bomze et al. "Annealed replication: A new heuristic for the maximum clique problem". In: *Discrete Appl. Math.* 121.13 (2002), pp. 27–49.

[4] I. M. Bomze et al. *The Maximum Clique Problem.* Boston: Springer, 1999, pp. 1–74.

[5] Immanuel M. Bomze. "Evolution towards the Maximum Clique". In: *Journal of Global Optimization* 10 (1997), pp. 143–164. URL: https://api.semanticscholar.org/CorpusID:33853863.

[6] Immanuel M. Bomze. "Regularity versus Degeneracy in Dynamics, Games, and Optimization: A Unified Approach to Different Aspects". In: *SIAM Review* 44.3 (2002), pp. 394–414. ISSN: 00361445. URL: http://www.jstor.org/stable/4148380 (visited on 08/23/2023).

[7] Immanuel M. Bomze, Francesco Rinaldi, and Samuel Rota Bulò. "First-order Methods for the Impatient: Support Identification in Finite Time with Convergent Frank–Wolfe Variants". In: *SIAM Journal on Optimization* 29.3 (2019), pp. 2211–2226. DOI: 10.1137/18M1206953. eprint: https://doi.org/10.1137/18M1206953. URL: https://doi.org/10.1137/18M1206953.

[8] Immanuel M. Bomze, Francesco Rinaldi, and Damiano Zeffiro. *Active set complexity of the Away-step Frank-Wolfe Algorithm.* 2019. arXiv: 1912.11492 [math.OC].

[9] Immanuel M. Bomze, Francesco Rinaldi, and Damiano Zeffiro. *Fast cluster detection in networks by first-order optimization.* 2021. arXiv: 2103.15907 [math.OC].

[10] Samuel Rota Bulò and Marcello Pelillo. "A generalization of the Motzkin–Straus theorem to hypergraphs". In: *Optimization Letters* 3 (2009), pp. 287–295. DOI: 10.1007/s11590-008-0108-3.

[11] Philip S. Chodrow, Nate Veldt, and Austin R. Benson. *Generative hypergraph clustering: from blockmodels to modularity.* 2021. arXiv: 2101.09611 [cs.SI].

[12] Marguerite Frank and Philip Wolfe. "An algorithm for quadratic programming". In: *Naval Research Logistics Quarterly* 3 (1956), pp. 95–110. URL: https://api.semanticscholar.org/CorpusID:122654717.

[13] Dan Garber. *Revisiting Frank-Wolfe for Polytopes: Strict Complementarity and Sparsity.* 2020. arXiv: 2006.00558 [math.OC].

[14] L. E. Gibbons et al. "Continuous characterizations of the maximum clique problem". In: *Math. Oper. Res.* 22.3 (1997), pp. 754–768.

[15] James T. Hungerford and Francesco Rinaldi. *A General Regularized Continuous Formulation for the Maximum Clique Problem.* 2017. arXiv: 1709.02486 [math.OC].

[16] D. S. Johnson. *Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge.* Vol. 26. DIMACS Ser. Discrete Math. Theoret. Comput. Sci. Providence, RI: American Mathematical Society, 1993, pp. 11–13.

[17] Jürgen Jost and Dong Zhang. *Discrete-to-Continuous Extensions: Lovász extension and Morse theory.* 2022. arXiv: 2003.06021 [math.CO].

[18] Jürgen Jost and Dong Zhang. *Discrete-to-Continuous Extensions: Lovász extension and Morse theory.* 2022. arXiv: 2003.06021 [math.CO].

[19] Jürgen Jost and Dong Zhang. *Discrete-to-Continuous Extensions: Lovász extension and Morse theory.* 2022. arXiv: 2003.06021 [math.CO].

[20] Jürgen Jost and Dong Zhang. *Discrete-to-Continuous Extensions: piecewise multilinear extension, min-max theory and spectral theory.* 2021. arXiv: 2106.04116 [math.CO].

[21] R. Karp. "Reducibility among combinatorial problems". In: (1972). Ed. by R. E. Miller and J. W. Thatcher, pp. 85–103.

[22] A. Kuznetsova and A. Strekalovsky. "On Solving the Maximum Clique Problem". In: *J. Global Optim.* 21.3 (2001), pp. 265–288.

[23] Chang Liang and Yuning Yang. "A note on Banach's results concerning homogeneous polynomials associated with nonnegative tensors". In: 15 (Mar. 2021), pp. 419–429. DOI: 10.1007/s11590-020-01602.

[24] Duncan Luce and Albert D. Perry. "A method of matrix analysis of group structure". In: *Psychometrika* 14 (1949), pp. 95–116. URL: https://api.semanticscholar.org/CorpusID:16186758.

[25] T. S. Motzkin and E. G. Straus. "Maxima for Graphs and a New Proof of a Theorem of Turán". In: *Canadian J. Math.* 17 (1965), pp. 533–540.

[26] Panos Pardalos and Jue Xue. "The Maximum Clique Problem". In: *Journal of Global Optimization* 4 (Apr. 1994), pp. 301–328. DOI: 10.1007/BF01098364.

[27] Bharath Pattabiraman et al. "Fast Algorithms for the Maximum Clique Problem on Massive Graphs with Applications to Overlapping Community Detection". In: *Internet Mathematics* 11.4-5 (May 2015), pp. 421–448. DOI: 10.1080/15427951.2014.986778. URL: https://doi.org/10.1080%2F15427951.2014.986778.

[28] Jeffrey Pattillo, Nataly Youssef, and Sergiy Butenko. "On clique relaxation models in network analysis". In: *European Journal of Operational Research* 226.1 (2013), pp. 9–18. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2012.10.021. URL: https://www.sciencedirect.com/science/article/pii/S0377221712007679.

[29] M. Pelillo. "Relaxation Labeling Networks for the Maximum Clique Problem". In: *J. Artificial Neural Networks* 2.4 (1996), pp. 313–328.

[30] M. Pelillo and A. Jagota. "Feasible and Infeasible Maxima in a Quadratic Program for Maximum Clique". In: *J. Artificial Neural Networks* 2.4 (1996), pp. 411–420.

[31] Francesco Rinaldi and Damiano Zeffiro. *Avoiding bad steps in Frank Wolfe variants*. 2022. arXiv: 2012.12737 [math.OC].

[32] S. Rota-Bulò and M. Pelillo. "A Generalization of the Motzkin-Straus Theorem to Hypergraphs". In: *Optim. Lett.* 3.2 (2009), pp. 287–295.

[33] Nicolò Ruggeri et al. *Community Detection in Large Hypergraphs*. 2023. arXiv: 2301.11226 [cs.SI].

[34] Vladimir Stozhkov et al. *Continuous cubic formulations for cluster detection problems in networks*. 2022. DOI: 10.1007/s10107-020-01572-4.

[35] Svyatoslav Trukhanov et al. "Algorithms for Detecting Optimal Hereditary Structures in Graphs, with Application to Clique Relaxations". In: *Comput. Optim. Appl.* 56.1 (Sept. 2013), pp. 113–130. ISSN: 0926-6003. DOI: 10.1007/s10589-013-9548-5. URL: https://doi.org/10.1007/s10589-013-9548-5.

[36] Sara Venturini et al. *Laplacian-based Semi-Supervised Learning in Multilayer Hypergraphs by Coordinate Descent*. 2023. arXiv: 2301.12184 [cs.LG].

[37] Q. Wu and J. K. Hao. "A review on algorithms for maximum clique problems". In: *Eur. J. Oper. Res.* 242.3 (2015), pp. 693–709.

[38] Qinghua Wu and Jin-Kao Hao. "A review on algorithms for maximum clique problems". In: *European Journal of Operational Research* 242 (May 2015), pp. 693–709. DOI: 10.1016/j.ejor.2014.09.064.

[39] Haiyuan Yu et al. "Predicting interactions in protein networks by completing defective cliques". In: *Bioinformatics* 22.7 (Feb. 2006), pp. 823–829. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl014. eprint: https://academic.oup.com/bioinformatics/article-pdf/22/7/823/48841534/bioinformatics_22_7_823.pdf. URL: https://doi.org/10.1093/bioinformatics/btl014.

[40] Yaoming Zhen and Junhui Wang. *Community Detection in General Hypergraph via Graph Embedding*. 2021. arXiv: 2103.15035 [stat.ML].