The Local Convexification Method and its Application to Learning Weakly Convex Boolean Functions

Eike Stadtländer[1,2], Tamás Horváth[1,2,3] (✉), and Stefan Wrobel[1,2,3]

[1] Dept. of Computer Science, University Bonn, Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany
[3] Fraunhofer IAIS, Sankt Augustin, Germany
{stadtlaender,horvath,wrobel}@cs.uni-bonn.de

Abstract. We study the problem of finding consistent hypotheses over finite metric spaces, focusing on hypothesis classes formed by weakly convex subsets of the domain. These hypotheses are closed under geodesics of length below a given threshold and exhibit a natural partitioning. Their generalization performance is strongly correlated with the number of blocks in the partition: fewer blocks yield greater generalization power. We prove that finding consistent weakly convex hypotheses with a minimum number of blocks is NP-hard. To address this negative result, we propose a novel greedy heuristic for computing compact solutions across a broad class of metric spaces and analyze its formal properties. Unlike standard approaches that calculate a single global distance threshold, our heuristic dynamically adjusts multiple local thresholds to seek compact hypotheses. To evaluate our method, we consider the specific case where the underlying metric space is the Hamming space, corresponding to learning weakly convex Boolean functions. Our empirical results demonstrate that our general-purpose algorithm outperforms the method specifically designed for learning this kind of Boolean functions in both model compactness and predictive performance. In fact, our approach generates hypotheses that are near-optimal with respect to the number of blocks in most cases.

Keywords: concept learning · consistent hypothesis finding · finite metric spaces · convexity · k-convex Boolean functions

1 Introduction

One of the core problems of supervised learning is the *consistent hypothesis finding* (CHF) problem. This problem involves identifying a hypothesis within a predefined hypothesis space that achieves zero empirical error on a given set of training examples. The CHF problem has been studied for numerous concept learning problems over different domains, including tabular data, propositional logic, first-order logic, graphs, and geometric concepts (see, e.g., [1, 2, 17, 18, 20–22, 24]). In this work we propose a *generic* heuristic that efficiently solves the CHF problem for a *broad* class of concept learning problems defined over finite *metric spaces*. Building on the fact that metric spaces admit the notion of weak convexity, our algorithm solves the CHF problem by computing consistent weakly convex hypotheses as in [14, 23]. Unlike [14, 23], however, our approach produces more compact hypotheses with better generalization performance. These hypotheses are subsets of the domain, consisting of pairwise disjoint blocks that are distant from one another. They are closed under geodesics between specific point pairs. More precisely, for any pair of points with a distance below a threshold *specific* to the pair, all points on all geodesics between the two points are included in the subset.

The idea of using weakly convex hypotheses was coined in [14] for learning *Boolean functions*. Specifically, [14] restricts the hypothesis space to Boolean functions with a particular property: Their true points (those satisfying the function) can be partitioned into subcubes of the d-dimensional Boolean cube such that each pair of subcubes has a Hamming distance greater than a positive integer k. Such Boolean functions can be can be represented by disjunctive normal forms (DNFs) in which the subcubes corresponding to the conjunctive terms are pairwise disjoint. One of the main results of [14] is that the CHF problem can be solved in *polynomial* time for hypothesis spaces formed by these so-called *k-convex* Boolean functions. The key insight is that, for some $k \geq 0$, there always exists a largest k-convex hull of the positive examples that is disjoint from the negative examples. Building on this, [23] generalizes the concept of k-convex Boolean functions and the associated CHF algorithm, extending them to weakly convex hypotheses over a broader class of metric spaces.

The generalization power of weakly convex hypotheses is strongly correlated with the number of their blocks. Specifically, hypotheses with fewer blocks exhibit higher generalization performance. This raises an important question: Can consistent weakly convex hypotheses with the *minimum* number of blocks be found in polynomial time? As a first contribution, we answer this question *negatively*, proving that the problem is NP-complete.

In line with this negative result, both algorithms in [14] and [23] fail to effectively optimize the number of blocks. A closer examination reveals that these approaches determine a *global* distance threshold for the output hypothesis. This threshold is defined as the minimum distance between two blocks across *all* block pairs (e.g., pairs of subcubes of the Boolean cube in the case of k-convex Boolean functions) that cannot be merged without violating consistency. However, the global nature of this threshold has a detrimental effect: it can force block pairs to remain separate, even if they could be merged into a single block at a larger distance threshold without violating consistency.

As a second contribution, we propose the LOCAL CONVEXIFICATION METHOD (LCM) to address this problem. It constructs *compact* weakly convex hypotheses by greedily merging blocks in order of increasing distance, while maintaining consistency and dynamically adjusting distance thresholds. A distinguishing feature of LCM, compared to [14, 23], is its ability to compute multiple *local* thresholds that vary across different pairs of points. Applying LCM requires implementing certain operations (e.g. the join operation for blocks) specific to the underlying metric space. To illustrate this, we consider k-convex Boolean functions as an example, demonstrating that this step generally does not present significant challenges. Importantly, we emphasize that LCM is a *general* method applicable to CHF problems across various finite metric spaces, not just the Hamming space.

We study some formal properties of LCM. In particular, we show that it is *sound* (i.e., it returns a consistent weakly convex hypothesis). Moreover, LCM is *efficient* whenever the abovementioned functions specific to the underlying metric space can be computed in polynomial time. Regarding *optimality* in terms of the number of blocks, the hypotheses generated by LCM are at least as compact as those produced by the algorithm in [23]. Additionally, LCM can achieve an exponential-size compression ratio compared to the hypotheses returned by [23]. As a further contribution, we conduct an experimental evaluation of LCM, our *general-purpose* heuristic, in the special case of Hamming spaces. We compare its performance against the CHF algorithm designed for k-convex Boolean functions in [14] and against DNFs extracted from Boolean decision trees. Our results demonstrate that LCM significantly outperforms both baselines in terms of model compactness and predictive performance. Furthermore, the number of terms in the DNFs produced by LCM is very close to the optimal value.

The rest of the paper is organized as follows. Section 1.1 reviews related work, while Section 2 introduces the necessary background notions. Section 3 defines locally constrained block systems, which constitute the hypotheses of the hypothesis class explored in this study. The negative result concerning the complexity of finding block-minimum consistent weakly convex hypotheses, along with the proposed heuristic, is presented in Section 4. The experimental results are presented and discussed in Section 5. Finally, Section 6 concludes the paper and suggests potential directions for future research.

Due to space limitations, we omit most proofs and offer a simplified adaptation of our approach–originally developed for *interval convexity* [4]–to the case of geodesic convexity. Full formal statements and their proofs will be provided in an extended version of this work.

1.1 Related Work

Closure systems (resp. closed sets) [11] can be regarded as a generalization of the family of all convex subsets of $\mathbb{R}^d$ (resp. convex sets in $\mathbb{R}^d$). Abstract closure systems have been studied also in the context of the CHF problem, e.g., in [17, 21]. Recently, there is an increasing interest in *geodesic* or *shortest-path* convexity [25] in machine learning, e.g., for vertex classification in graphs [12, 22–24] and recovering clusterings [3].

The relaxation of convexity to *weak convexity* and to similar notions was studied before for *discrete* metric spaces (see, e.g., [6–9]). In machine learning, k-convex Boolean functions were first investigated in [14]. In our general results, we utilize abstract interval functions [4] to extend the concept of weak convexity to weak interval convexity. Our notion of *locally constrained block systems* generalizes the concept of θ-decompositions of *weakly convex sets* defined in [23]. In particular, the CHF problem for learning weakly convex sets is solved in [23] (see, also, [14]) by finding the largest θ-convex *hull* of the positive examples over all $\theta \geq 0$ that is consistent with the negative examples. We are interested in consistent locally constrained block systems, i.e., consistent hypotheses formed by unions of weakly (interval) convex sets, that are more compact in terms of their number of blocks. This is due to the property that the number of blocks is inversely correlated with the block system's generalization power.

2 Preliminaries

This section collects the necessary concepts and defines the notation. Unless otherwise stated, all sets and metric spaces are assumed to be *finite*. Special attention will be given to the *Hamming space*, denoted by $\mathcal{M}_H = (\mathbf{B}_d, D_H)$, where $\mathbf{B}_d = \{0,1\}^d$ is the d-dimensional Boolean cube and D_H denotes the Hamming distance.

The *power set* of a set X is denoted by 2^X. A *closure system* (see, e.g., [11]) $\mathcal{C}$ over a set X is a collection of subsets of X that contains X and is closed under arbitrary intersections. It has a corresponding *closure operator* $\rho : 2^X \to 2^X$ that is *extensive* (i.e., $A \subseteq \rho(A)$ for all $A \in 2^X$), *monotone* (i.e., $\rho(A) \subseteq \rho(B)$ for all $A, B \in 2^X$ with $A \subseteq B$), and *idempotent* (i.e., $\rho(\rho(A)) = \rho(A)$ for all $A \in 2^X$) and which satisfies $C \in \mathcal{C}$ iff $C = \rho(C)$ for all $C \subseteq X$. The fixpoints of ρ are referred to as *closed* sets.

Ordinary and abstract *convexity* are used explicitly or implicitly by many learning algorithms. Examples include support vector machines [10] (half-spaces are convex subsets of $\mathbb{R}^d$) or learning separating half-spaces in graphs [22] (half-spaces are defined by abstract convexity over the vertex set of a graph). For a metric space $\mathcal{M} = (X, D)$, $C \subseteq X$ is *geodesically convex* or simply, *convex* if for all $x, y \in C$ and $z \in X$, $z \in C$ whenever z lies on a geodesic between x and y, i.e., $D(x,y) = D(x,z) + D(z,y)$. The family of all convex subsets of a metric space $\mathcal{M} = (X, D)$ is denoted by $\mathcal{C}$. As an example, consider the function over $\mathbf{B}_d$ that maps all subsets A of $\mathbf{B}_d$ to the smallest subcube C of $\mathbf{B}_d$ that contains A. It is elementary to check that C is convex for $\mathcal{M}_H$. Note that C can be represented by a conjunction over $2d$ Boolean literals.

Hypothesis classes formed by convex subsets of the domain can be disadvantageous for machine learning, as they cannot capture *multiple well-separated* regions of interest. This limitation is addressed in [23] by generalizing the notion of convexity to that of *weak convexity*: A subset C of a metric space (X, D) is *weakly convex* (or θ*-convex*) for some $\theta \geq 0$ if, for all $x, y \in C$ and $z \in X$, $z \in C$ whenever $D(x,y) \leq \theta$ and $D(x,y) = D(x,z) + D(z,y)$. The collection of all θ-convex subsets of X is denoted by $\mathcal{C}_\theta$.

Our focus will be on hypotheses formed by pairwise disjoint "contiguous" blocks. To this end, we need to define the notion of "contiguity" for metric spaces. Specifically, a sequence $x_1, \ldots, x_\ell$ of pairwise distinct elements of X forms a θ*-path* for some $\theta \geq 0$, if $D(x_i, x_{i+1}) \leq \theta$ for all $i = 1, \ldots, \ell - 1$. A set $A \subseteq X$ is θ*-connected* if for all $x, y \in A$, there exists a θ-path in A connecting x and y (i.e., $x_1 = x$ and $x_\ell = y$). Clearly, A is $\mathrm{diam}(A)$-connected, where $\mathrm{diam}(A) = \max\{D(x,y) : x, y \in A\}$ is the *diameter* of A. The proof of the proposition below is immediate from the definitions.

Proposition 1. *Let (X, D) be a metric space and $A \subseteq X$. Then for all $\theta \geq 0$,*

(i) if A is θ-connected then A is θ'-connected for all $\theta' \geq \theta$,
(ii) A has a unique θ-partitioning defined by θ-connectivity, i.e., for all $x, y \in A$, x and y are in the same θ-connected component if and only if they are θ-connected.

The *connectivity index* of A in the above proposition, denoted by $\mathrm{CI}(A)$, is defined as the smallest value θ for which A is θ-connected. This is well-defined, as X is finite.

Theorem 1 below, a decomposition result from [23], provides a characterization of θ-convex sets. Specifically, it states that $\mathcal{C}_\theta$, the collection of θ-convex sets, forms a closure system. Moreover, every θ-convex set can be expressed as a family of θ-*connected* and θ-*convex* sets that are pairwise θ-distant from each other.

Theorem 1. *Let $\theta \geq 0$ and $\mathcal{M} = (X, D)$ be a metric space. Then (i) $\mathcal{C}_\theta$ forms a closure system and (ii) for all $C \subseteq X$, $C \in \mathcal{C}_\theta$ if and only if there exists a family $\mathcal{P} = \{B_j\}_{j \in J}$*

for some index set J with $C = \bigcup_{j \in J} B_j$ that satisfies the following properties for all $j \in J$:

(α) B_j is θ-convex,
(β) B_j is θ-connected, and
(γ) for all $i \in J$ with $i \neq j$, $D(B_i, B_j) > \theta$.

As a consequence of Proposition 1, $\mathcal{P}$ in the theorem above forms a *unique* partition of C. We will therefore refer to $\mathcal{P}$ as the *θ-convex decomposition* of C. The B_js will be called *θ-convex blocks*, or simply *blocks* of C. Furthermore, (i) of Theorem 1 implies that there exists a *closure operator* $\rho_\theta : 2^X \to 2^X$ with $A \mapsto \bigcap\{C \in \mathcal{C}_\theta : A \subseteq C\}$ for all $A \subseteq X$. In other words, ρ_θ maps A to the (unique) smallest θ-convex set in $\mathcal{C}_\theta$ that contains A. Henceforth, it will be referred to as the *θ-convex hull* of A.

The fundamental properties of weak convexity, as stated in Lemma 1 below, will be used frequently throughout the remainder of this paper.

Lemma 1. *Let $\mathcal{M} = (X, D)$ be a metric space, $\theta \geq \theta' \geq 0$, and $S \subseteq X$. Then*

(i) $\mathcal{C} \subseteq \mathcal{C}_\theta \subseteq \mathcal{C}_{\theta'}$, i.e., convexity implies θ-convexity, which in turn implies θ'-convexity,
(ii) $\rho_{\theta'}(S) \subseteq \rho_\theta(S)$, and
(iii) if $C = \rho_\theta(S)$ and $\mathcal{P} = \{B_j\}_{j \in J}$ is the θ-convex decomposition of C, then $B_j = \rho_\theta(S \cap B_j)$ for all $j \in J$.

Properties (i) and (ii) of Lemma 1 establish that monotonicity for the θ-convex hulls holds not only with respect to the input set but also to the distance threshold θ. This result, when combined with the decomposition theorem (Theorem 1), implies that the generators of a θ-convex set C determine its blocks, as stated in Property (iii).

3 Locally Constrained Block Systems

This paper is concerned with the *consistent hypothesis finding* (CHF) problem for learning weakly convex concepts. We first consider the following CHF problem:

Problem 1. Given a metric space $\mathcal{M} = (X, D)$ and $E^+, E^- \subseteq X$, *find* a $\theta \geq 0$ and a θ-convex set $H \in \mathcal{C}_\theta$ that is *consistent* with E^+ and E^- (i.e., $E^+ \subseteq H$ and $E^- \cap H = \emptyset$); or return "NO" if there is no such H.

To solve Problem 1, [23] (cf. [14]) employs the decomposition theorem (Theorem 1). Specifically, the solution involves computing the *largest* θ-convex hull of E^+ over all $\theta \geq 0$ that remains disjoint from E^-. Since $\rho_0(A) = A$ for all $A \subseteq X$, a consistent hypothesis is guaranteed to exist when E^+ and E^- are disjoint. Depending on the context, this hypothesis will be referred to as the *consistent globally constrained θ-convex hypothesis*, or simply the *consistent θ-*GC *hypothesis*.

Example 1. To illustrate the above concepts, consider the Hamming space $\mathcal{M}_H = (\mathbf{B}_d, D_H)$ for $d = 8$. Let $E^+ = \{u_1, u_2, v_1, \ldots, v_4\}$ and $E^- = \{w\}$ with

$$\begin{aligned} &u_1 = (00001111), \quad u_2 = (00010111) \\ &v_1 = (11111000), \quad v_2 = (11111011), v_3 = (11111101), v_4 = (11111110) \\ &w = (00100111)\ . \end{aligned}$$

The consistent θ-GC hypothesis is attained for $\theta = 2$. To seee this, note first that $\rho_2(E^+)$, which can be represented by the DNF $\phi = \overline{x}_1\overline{x}_2\overline{x}_3x_6x_7x_8 \vee x_1x_2x_3x_4x_5$, is consistent, as it is not satisfied by the negative example w. Furthermore, $w \in \rho_\theta(E^+) = \mathbf{B}_8$ for all $\theta > 2$. Notice that the two subcubes represented by the terms in ϕ fulfill all properties required in (ii) of Theorem 1 for the blocks of a 2-convex decomposition of $\rho_\theta(E^+)$. Notably, they are separated by a distance of 3 from each other. □

A major limitation of the approach in [23] is that the consistent θ-GC hypothesis may contain an excessively high number of blocks. This can lead to an overly specific solution, potentially resulting in overfitting. The issue arises from a fundamental property of the approach: Although the value of θ corresponding to the θ-GC hypothesis is determined by *local* regions induced by the training examples, it is applied *globally* across the entire training set. The following example illustrates this problem.

Example 2. Consider the set E^+ of positive examples in Example 1. Let $E^- = \{w'\}$, where $w' = (00011111)$. Since the negative example w' satisfies the conjunction $\overline{x}_1\overline{x}_2\overline{x}_3x_6x_7x_8$, which represents $\rho_2(\{u_1, u_2\})$, we have $w' \in \rho_2(\{u_1, u_2\}) \subset \rho_2(E^+)$. Consequently, given that $\rho_0(E^+) = \rho_1(E^+) = E^+$, the consistent θ-GC hypothesis is identical for $\theta = 0$ and $\theta = 1$, comprising $|E^+|$ singleton blocks. The local distance constraint defined by u_1, u_2, and w' prevents the algorithm from generalizing the positive examples $v_1, \ldots, v_4$ for all $\theta > 1$. However, they can be generalized for $\theta = 2$ without violating consistency. Specifically, $\rho_2(\{v_1, \ldots, v_4\})$, represented by $x_1x_2x_3x_4x_5$, does not contain w' and is at a distance of $4 > 2$ from both u_1 and u_2. □

To address this limitation, we consider other weakly convex sets as potential candidate hypotheses. The above observations motivate the following definition.

Definition 1 (Locally Constrained Block Systems). *Let $\mathcal{M} = (X, D)$ be a metric space and $\theta \geq 0$. A set $\mathcal{B} = \{(B_j, \theta_j)\}_{j \in J}$ with $\theta_j \geq \theta$ for some index set J is a* locally constrained block system *for θ, or θ-*LC *block system for short, if the following properties hold for all $j \in J$:*

(α') B_j is θ_j-convex,
(β') B_j is θ_j-connected,
(γ') for all $i \in J$ with $i \neq j$, $D(B_i, B_j) > \max\{\theta_i, \theta_j\}$.

A set $A \subseteq X$ is *covered* by $\mathcal{B}$ in Definition 1 if $A \subseteq \operatorname{dom}(\mathcal{B})$, where $\operatorname{dom}(\mathcal{B}) = \bigcup_{j \in J} B_j$ denotes the *domain* of $\mathcal{B}$. Definition 1 is inspired by the characterization of weakly convex sets in (ii) of Theorem 1. The key distinctions between (α) and (α') and between (β) and (β') lie in the *relaxation* of the distance thresholds from the global θ in Theorem 1 to some local $\theta_j \geq \theta$. Consequently, the pairwise distance constraints in (γ) must hold for the maximum of the blocks' distance thresholds in (γ'). It is worth noting that the θ_j values in the definition are not required to be pairwise distinct. Furthermore, different θ-LC block systems can share the same domain.

Example 3. For E^+ in Example 2 we have that $\mathcal{B} = \{(T_1, 0), (T_2, 0), (T_3, 2)\}$ is a 0-LC block system, where $T_1 = \overline{x}_1\overline{x}_2\overline{x}_3\overline{x}_4x_5x_6x_7x_8, T_2 = \overline{x}_1\overline{x}_2\overline{x}_3x_4\overline{x}_5x_6x_7x_8$, and $T_3 = x_1x_2x_3x_4x_5$ represent $\rho_0(\{u_1\}) = \{u_1\}, \rho_0(\{u_2\}) = \{u_2\}$, and $\rho_2(\{v_1, \ldots, v_4\})$, repectively. □

We note that $\mathrm{dom}(\mathcal{B})$ of a θ-LC block system $\mathcal{B}$ is the union of pairwise distant weakly convex sets, each discretely connected for some (local) distance threshold (see Example 3). The decomposition result in Theorem 1 suggests that $\mathrm{dom}(\mathcal{B})$ itself is a weakly convex set. Proposition 2 below addresses this observation (cf. (ii)). However, although each block B_j of $\mathcal{B}$ is θ_j-connected (cf. (iii)), the θ-convex decomposition of $\mathrm{dom}(\mathcal{B})$ might comprise even more blocks, as θ_j may be strictly larger than θ.

Proposition 2. *Let $\mathcal{M} = (X, D)$ be a metric space, $\theta \geq 0$, and $\mathcal{B} = \{(B_j, \theta_j)\}_{j \in J}$ a θ-*LC *block system over $\mathcal{M}$. Then*

(i) $\mathrm{dom}(\mathcal{B}) \subseteq X$*,*
(ii) $\mathrm{dom}(\mathcal{B})$ *is θ-convex,*
(iii) for all $j \in J$, B_j is a θ_j-convex block, and
*(iv) for every $J' \subseteq J$ with $J' \neq \emptyset$, $\mathcal{B}' = \{(B_j, \theta_j)\}_{j \in J'}$ is a θ-*LC *block system.*

Conversely, Proposition 4 below asserts that the θ-convex decomposition of a θ-convex set $C \subseteq X$ can be regarded as a θ-LC block system in a straightforward manner. Moreover, this *canonical* θ-LC block system is the “finest” among all possible θ-LC block systems covering C, in the following sense: A θ-LC block system $\mathcal{B} = \{(B_j, \theta_j)\}_{j \in J}$ over a metric space $\mathcal{M} = (X, D)$ for some $\theta \geq 0$ is considered *coarser* than a θ-LC block system $\mathcal{B}' = \{(B'_k, \theta'_k)\}_{k \in K}$ for some $\theta' \geq 0$, denoted $\mathcal{B} \preccurlyeq \mathcal{B}'$ (or equivalently, $\mathcal{B}'$ is *finer* than $\mathcal{B}$, denoted $\mathcal{B}' \succcurlyeq \mathcal{B}$), if for every $k \in K$ there exists $j \in J$ such that $B'_k \subseteq B_j$. Clearly, $\mathcal{B} \preccurlyeq \mathcal{B}'$ implies $\mathrm{dom}(\mathcal{B}) \supseteq \mathrm{dom}(\mathcal{B}')$. Our focus will be on θ-LC block systems that contain *no* irrelevant blocks with respect to the set of positive examples, meaning that every block contains at least one positive example. Specifically, $\mathcal{B}$ is S*-relevant* for some $S \subseteq X$ if $S \subseteq \mathrm{dom}(\mathcal{B})$ and for every $j \in J$, $S \cap B_j \neq \emptyset$. We restrict the $\preccurlyeq$ relation to this type of block systems. In particular, $\mathcal{B} \preccurlyeq_S \mathcal{B}'$ denotes that $\mathcal{B} \preccurlyeq \mathcal{B}'$ and both $\mathcal{B}$ and $\mathcal{B}'$ are S-relevant. As mentioned earlier, S will later be restricted to the set of positive examples. Clearly, $\mathcal{B} \preccurlyeq_S \mathcal{B}'$ implies that all blocks of $\mathcal{B}$ contain a block of $\mathcal{B}'$, leading to the following claim.

Proposition 3. *Let $\mathcal{M} = (X, D)$ be a metric space, $S \subseteq X$, and let $\mathcal{B} = \{(B_j, \theta_j)\}_{j \in J}$ and $\mathcal{B}' = \{(B_k, \theta_k)\}_{k \in K}$ be θ-*LC *and θ'-*LC *block systems, respectively, for some $\theta, \theta' \geq 0$. If $\mathcal{B} \preccurlyeq_S \mathcal{B}'$, then $|J| \leq |K|$, i.e., the number of blocks in $\mathcal{B}$ is bounded by that in $\mathcal{B}'$.*

We employ the notation $\succ_S$ and $\prec_S$ when $|J| < |K|$. In the following proposition, we establish a relationship between the θ-convex decomposition of a θ-convex set $C \subseteq X$ and the θ-LC block systems that cover C.

Proposition 4. *Let $\mathcal{M} = (X, D)$ be a metric space, $\theta \geq 0$, $\mathcal{B} = \{(B_i, \theta_i)\}_{j \in J}$ a θ-*LC *block system, $C \subseteq X$ a θ-convex set covered by $\mathcal{B}$, and $\mathcal{P} = \{P_k\}_{k \in K}$ the θ-convex decomposition of C. Then*

*(i) $\mathcal{B}' = \{(P_k, \theta)\}_{k \in K}$ is a θ-*LC *block system with* $\mathrm{dom}(\mathcal{B}') = C$*,*
*(ii) $\mathcal{B} \preccurlyeq \mathcal{B}'$, i.e., any θ-*LC *block system covering C is coarser than the θ-convex decomposition of C, and*
*(iii) if $\mathcal{B}$ is C-relevant then $|J| \leq |K|$, i.e., the number of blocks in a C-relevant θ-*LC *block system is bounded by that in the θ-convex decomposition of C.*

Proposition 4 (iii) suggests that θ-LC block systems can cover the consistent θ-GC hypothesis studied in [23], potentially using fewer blocks. In Section 5, we provide experimental evidence demonstrating that the consideration of this broader hypothesis class for the CHF problem results in a substantial reduction in the number of blocks.

4 The Local Convexification Method

Under some natural assumptions, the consistent θ-GC hypothesis can be found in polynomial time [23]. As discussed earlier, it is not optimal in terms of the number of blocks compared to consistent θ-LC block systems; a θ-LC block system $\mathcal{B}$ is *consistent* with the sets E^+ and E^- of positive and negative examples if $E^+ \subseteq \mathrm{dom}(\mathcal{B})$ and $E^- \cap \mathrm{dom}(\mathcal{B}) = \emptyset$. The following theorem demonstrates, in the specific case of k-convex Boolean functions, that a consistent θ-LC block system can be *exponentially* more compact than the consistent θ-GC block system considered in [14, 23].

Theorem 2. *For all sufficiently large positive integers d, there exist $E^+, E^- \subseteq \mathbf{B}_d$ and a θ-*LC *block system $\mathcal{B}$ consistent with E^+ and E^- such that the size of the consistent θ-*GC *block system $\mathcal{B}_c$ relative to $\mathcal{B}$ satisfies*

$$\frac{|\mathcal{B}_c|}{|\mathcal{B}|} = 2^{\Omega(d)} \quad .$$

Proof. Let $d' = d - 7$ and let S be a largest subset of $\mathbf{B}_{d'}$ such that the pairwise Hamming distance between any two elements of S is at least 3. By the Gilbert-Varshamov bound (see, e.g., Chapter 8 in [16]), a fundamental result in coding theory, we have

$$|S| \geq \frac{2^{d'}}{\sum_{j=0}^{2} \binom{d'}{j}} = 2^{\Omega(d)} \quad . \tag{1}$$

Let $E^+ = \{x, y\} \cup S'$ and $E^- = \{z\}$, where $x = 0^3 0^4 0^{d-7}$, $y = 1^3 0^4 0^{d-7}$, $z = 0010^4 0^{d-7}$, and $S' = \{0^3 1^4 \oplus s : s \in S\}$. Here, a^ℓ and $\oplus$ denote the ℓ-fold repetition of the symbol a and the string concatenation, respectively.

Since $D_H(x, y) = D_H(x, z) + D_H(z, y)$ and $D_H(x, y) = 3$, there is no consistent 3-convex Boolean function. In contrast, there exists a consistent hypothesis for $\theta = 2$. On the one hand, the consistent globally constrained θ-convex hypothesis $\mathcal{B}_c$ for $\theta = 2$ contains $|S| + 2 = 2^{\Omega(d)}$ blocks by (1), each of which is a singleton. On the other hand, utilizing the fact that the convex hull of a set of at least 3 points of $\mathbf{B}_d$ is always 2-connected [13], the 2-LC block system $\mathcal{B} = \{(\{x\}, 2), (\{y\}, 2), (\text{convex hull of } S, 2)\}$ is consistent and contains only three blocks. □

Using the notion of optimality in terms of number of blocks, Theorem 2 gives rise to the following CHF problem.

Problem 2 (Block-Minimum CHF Problem). Given a metric space $\mathcal{M} = (X, D)$ and $E^+, E^- \subseteq X$, *find* a $\theta \geq 0$ and a consistent θ-LC block system $\mathcal{B} = \{(B_i, \theta_i)\}_{i \in \{1,\ldots,k\}}$ with the *smallest* k, or return "NO" if there are no such θ and $\mathcal{B}$.

The following theorem presents a negative result on the complexity of Problem 2.

Theorem 3. *Problem 2 is NP-complete.*

Proof. Given a finite set $\mathcal{B} \subseteq 2^X \times [0, \infty)$ over a metric space $\mathcal{M} = (X, D)$ and $\theta \geq 0$, it can be decided in polynomial time whether $\mathcal{B}$ is a θ-LC block system that is consistent with E^+ and E^-. Thus, Problem 2 is in NP. To show that it is NP-hard, we use a reduction from the disjoint version of the *boxes class cover* (BCC) problem [19] defined as follows: Given disjoint finite sets B of blue and R of red points in the plane, find a *minimum cardinality* set $\mathcal{H}$ of pairwise disjoint axis-aligned rectangles such that every blue point is contained in a rectangle and none of the red points belongs to any of the rectangles in $\mathcal{H}$. This problem is NP-complete [19, Theorem 4.10].

The main idea behind the reduction is that for an instance B, R of the disjoint version of the BCC problem, we construct a finite rectangular grid graph $G = (V, E)$ with vertices containing $B \cup R$ such that the shortest path between any two vertices in $B \cup R$ is at least 3, where the distance between the vertices is defined by the shortest-path distance. We have that a subset C of V is θ-convex for all $\theta \geq 2$ iff the subgraph of G induced by C is a grid graph. This property allows us to establish the connection between solutions of the disjoint version of the BCC problem containing k rectangles and those of Problem 2 containing k blocks.

More precisely, for an instance B, R of the disjoint version of the BCC problem with $|B \cup R| = n$, construct a graph $G = (V, E)$ as follows: For all $p \in B \cup R$, take a vertical and a horizontal line through p. Sort the points in $B \cup R$ according to their x-coordinates and for each adjacent points $(x, y), (x', y')$ with $x < x'$, select three values x_1, x_2, x_3 satisfying $x < x_1 < x_2 < x_3 < x'$ and take the three vertical lines through $(x_1, 0), (x_2, 0), (x_3, 0)$, respectively. In a similar way, sort the points in $B \cup R$ according to their y-coordinates and for each adjacent points with different y-coordinates take three pairwise different horizontal lines. In this way we obtain a grid in the plane with vertices defined by the set of pairwise intersections of the horizontal and vertical lines. Define V by the set of vertices of this grid and add edge $\{u, v\}$ to E iff they are adjacent in the grid. For any $u, v \in V$, define their distance $D(u, v)$ by their shortest-path distance in G. By construction, G is a rectangular grid graph with $B \cup R \subseteq V$ and $D(u, v) > 2$ for all $u, v \in R \cup B$ with $u \neq v$.

It holds that if $C \subseteq V$ induces a connected subgraph of G and C is θ-convex for some $\theta \geq 2$ then C is θ-convex for *all* $\theta \geq 2$ (i.e., C is convex) and that C induces a rectangular subgrid graph of G. Since the size of G is $\mathcal{O}(n^2)$, the reduction is polynomial.

It is easy to check that for all $k > 0$, there exists a hypothesis $\mathcal{H}$ containing k pairwise disjoint axis-aligned rectangles in $\mathbb{R}^2$ that is consistent with $B \cup R$ iff there exists a θ-LC block system $\mathcal{B}$ over $\mathcal{M}$ for $\theta = 2$ that consists of k blocks and is consistent with $E^+ = B$ and $E^- = R$. This completes the proof of the NP-hardness. □

4.1 The Algorithm

Motivated by the aforementioned negative result, we present a *greedy heuristic* called the LOCAL CONVEXIFICATION METHOD (LCM; see Algorithm 1) for finding *compact*

Algorithm 1 Local Convexification Method

```
Require: metric space $\mathcal{M} = (X, D)$ and threshold $\tau \geq 0$
Input: $E^+, E^- \subseteq X$
Output: $\theta$-LC block system $\mathcal{B}$ for some $\theta \geq \tau$ which is consistent with $E^+$ and $E^-$ if such $\theta$
    and $\mathcal{B}$ exist; otherwise “No”

$\mathcal{B} \leftarrow \{(B, \tau) : B \in \textsc{WeaklyConvexHull}(\tau, E^+)\},\ \mathcal{F} \leftarrow \emptyset$
if $\exists (B, \theta_B) \in \mathcal{B}, e \in E^-$ such that $\textsc{Membership}(e, B) = \textsc{True}$ then return “No”
while $\mathcal{A} := \{((R, \theta_R), (S, \theta_S)) \in \mathcal{B}^2 : R \neq S, \{R, S\} \notin \mathcal{F}\} \neq \emptyset$ do
    $\lambda \leftarrow \min\{\textsc{Distance}(R, S) : ((R, \theta_R), (S, \theta_S)) \in \mathcal{A}\}$
    choose $((R, \theta_R), (S, \theta_S)) \in \mathcal{A}$ such that $\textsc{Distance}(R, S) = \lambda$
    $\mathcal{D} \leftarrow \{(R, \theta_R), (S, \theta_S)\}$
    $B \leftarrow \textsc{Join}(\max\{\tau, \lambda\}, R, S),\ \ \theta_B \leftarrow \max\{\tau, \textsc{ConnectivityIndex}(B)\}$
    while $\exists (Q, \theta_Q) \in \mathcal{B} \setminus \mathcal{D}$ with $\textsc{Distance}(B, Q) \leq \max\{\tau, \theta_B, \theta_Q\}$ do
        $B \leftarrow \textsc{Join}(\max\{\tau, \theta_B, \theta_Q\}, B, Q), \mathcal{D} \leftarrow \mathcal{D} \cup \{(Q, \theta_Q)\}$
        $\theta_B \leftarrow \max\{\tau, \textsc{ConnectivityIndex}(B)\}$
    if $\exists e \in E^-$ such that $\textsc{Membership}(e, B) = \textsc{True}$ then $\mathcal{F} \leftarrow \mathcal{F} \cup \{\{R, S\}\}$
    else $\mathcal{B} \leftarrow (\mathcal{B} \setminus \mathcal{D}) \cup \{(B, \theta_B)\}$
return $\mathcal{B}$
```

consistent θ-LC block systems. The main idea of Algorithm 1 is to greedily *join* pairs of blocks in *ascending* order of their distance until any further join would result in an inconsistency.

Algorithm 1 operates on a finite metric space $\mathcal{M}$ and requires a threshold $\tau \geq 0$. Regarding τ, a user-specified lower bound on θ for the consistent θ-LC block system computed by the algorithm, it has been shown in [23] that certain metric spaces permit a *compact representation* of the blocks in θ-convex decompositions of θ-convex sets. In particular, some representation schemes make use of the fact that blocks are *convex* sets. For instance, the terms of a DNF representing a weakly convex Boolean function correspond to convex subsets of $\mathbf{B}_d$. This property can also be leveraged for θ-LC block systems. However, this requires the value of θ for a block to exceed a certain threshold τ, which is *intrinsically* tied to the underlying metric space. More precisely, a metric space is *blockwise convex* for some $\theta \geq 0$ if every θ-convex block (i.e., θ-connected and θ-convex set) is convex. For representation languages restricted to convex blocks, τ should be at least the smallest value of θ for which the metric space is blockwise convex. As an example, k-convex Boolean functions can be represented by DNFs whose conjunctive terms correspond precisely to the blocks in their respective decompositions [14]. In this specific case, $\tau = 2$, since the blocks of 2-convex subsets of $\mathcal{M}_H$ are convex [13].

In each iteration of the outer loop, Algorithm 1 computes a strictly coarser consistent θ-LC block system from the current consistent θ-LC block system $\mathcal{B}$ by joining block pairs in their increasing distance order. This is an iterated process that is repeated as long as the resulting block is consistent with the negative examples and does not violate any of the blocks' local distance constraints. If a join operation results in an inconsistency, the algorithm does *not* stop the computation. Instead, it returns to the state before the invalid join, adds the pair of blocks that led to the inconsistency to a set of *forbidden* pairs, and continues with the next pair of blocks.

We do not provide the pseudocode of the *subroutines* in Algorithm 1. Semantically, they are defined as follows: For $E^+ \subseteq X$, $e \in X$, blocks $B, R \subseteq X$, and $\theta \geq \tau$,

- WEAKLYCONVEXHULL(θ, E^+) computes the set of blocks of $\rho_\theta(E^+)$.
- MEMBERSHIP(e, B) decides whether or not $e \in B$.
- DISTANCE(B, R) computes the distance $D(B, R) = \min_{b \in B, r \in R} D(b, r)$.
- CONNECTIVITYINDEX(B) calculates the connectivity index $\mathrm{CI}(B)$ of B (i.e., the smallest θ' such that B is θ'-connected).
- JOIN(θ, B, R) computes the join of B and R defined by $\rho_\theta(B \cup R)$. Note that $\rho_\theta(B \cup R)$ constitutes a single block whenever $\theta \geq D(B, R)$, a condition that is always fulfilled during the entire execution of Algorithm 1.

Algorithm 1 starts by calling WEAKLYCONVEXHULL(τ, E^+) and initializing $\mathcal{F}$ as an empty set. It is used to store forbidden pairs of blocks, i.e., which cannot be joined. According to Proposition 4 (ii), this initial step produces the finest θ-LC block system containing E^+ for some $\theta \geq \tau$. If the resulting hypothesis is inconsistent with the negative examples E^-, the algorithm must return "NO", as there is no consistent hypothesis. This consistency check is performed in line 2. It follows from the definition of weakly convex hulls and Lemma 1 (iii) that the initial hypothesis $\mathcal{B}$ in line 1 is E^+-relevant.

When entering the main loop of Algorithm 1 (line 3), the properties of θ-LC block systems (see Definition 1) are preserved. Among the block pairs in $\mathcal{B}$ that are not in $\mathcal{F}$, a pair with the smallest distance, denoted λ, is selected (lines 4-5). However, the new block B, obtained by joining the two selected blocks (line 7), may violate local distance constraints with other blocks in relation to the updated connectivity index. To address this, further joins of violating blocks with B are computed in the inner loop (lines 8-10), if necessary. During this process, the connectivity index is updated to ensure that the properties of θ-LC block systems remain satisfied. Additionally, the join operation could potentially cause an inconsistency with E^-; this is checked in line 11. If an inconsistency is detected, the initial pair of blocks is added to the set $\mathcal{F}$ of forbidden joins. Otherwise, the data structure $\mathcal{B}$ is updated by removing the blocks that were joined into B and by adding B, along with the maximum of τ and its connectivity index θ_B. In each iteration of the main loop that modifies $\mathcal{B}$, the update results in a consistent *coarsening* of $\mathcal{B}$. This guarantees a consistent hypothesis for both E^+ and E^- while ensuring that it remains E^+-relevant. The following theorem addresses the soundness and computational complexity of Algorithm 1.

Theorem 4. *The following properties hold for Algorithm 1:*

*(i) It returns a consistent E^+-relevant θ-*LC *block system $\mathcal{B}$ for some $\theta \geq \tau$, or "*NO*" if such θ and $\mathcal{B}$ do not exist.*

(ii) It runs in time polynomial in $|E^+|$, $|E^-|$, and the parameters of the underlying metric space $\mathcal{M}$, provided that all five functions called by the algorithm also run in time polynomial in $|E^+|$ and the parameters of $\mathcal{M}$.

It remains to ask whether Algorithm 1 *always* returns at least a *block-minimal consistent* θ-LC block system. Specifically, a θ-LC block system $\mathcal{B}$ with $\theta \geq \tau$ is considered block-minimal consistent if it is consistent with E^+ and E^-, E^+-relevant, and

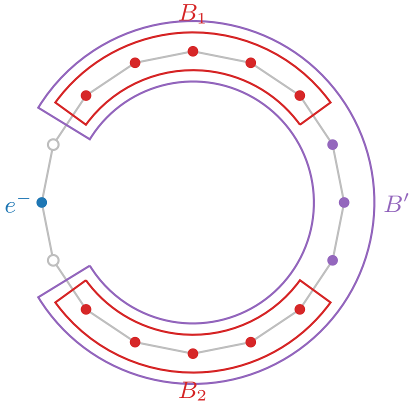

Fig. 1. An example for which Algorithm 1 returns a consistent θ-LC block system that is *not* block-minimal consistent. The output is $\mathcal{B} = \{(B_1, 1), (B_2, 1)\}$. However, $\mathcal{B} = \{(B', 1)\}$ is a consistent E^+-relevant 1-LC block system with $\mathcal{B} \succ_{E^+} \mathcal{B}'$.

there exists no consistent E^+-relevant θ'-LC block system $\mathcal{B}'$ for some $\theta' \geq \tau$ such that $\mathcal{B}' \prec_{E^+} \mathcal{B}$. Recall that E^+-relevance guarantees that $\mathcal{B}$ has no block that is disjoint with E^+. Note further that block-minimal consistency of a θ-LC block system does not imply that its number of blocks is minimum across all consistent θ-LC block systems. Example 4 below demonstrates that the answer to the above question is *negative*.

Example 4. We present an example which shows that the output of Algorithm 1 is *not* block-minimal consistent in general. The underlying metric space in this example is formed by the vertex set of an (unweighted) connected graph and the shortest-path distance. To this end, note first that $\mathrm{CI}(B) = 1$ for every $\theta \geq 1$ and for every θ-convex block B. Indeed, if B is a θ-convex block and $u, v \in B$, then there is a θ-path $u = p_1, p_2, \ldots, p_\ell = v$ between u and v that lies in B. Furthermore, since $D(p_i, p_{i+1}) \leq \theta$, all shortest paths between p_i and p_{i+1} are contained in B, for all $i = 1, \ldots, \ell - 1$. Choose one such shortest path between p_i and p_{i+1}, for all$i$. The concatenation of these paths is a 1-path between u and v. Hence, B is 1-connected implying $\mathrm{CI}(B) \leq 1$. Since $\mathrm{CI}(B) \geq 1$, $\mathrm{CI}(B) = 1$.

For the example, let $G = (V, E)$ denote the cycle consisting of 16 vertices given in Figure 1. Let the positive examples E^+ consist of the points depicted in **red** and the negative example $E^- = \{e^-\}$ be the single point depicted in **blue**. Consider the subgraphs B_1 and B_2 consisting of 5 positive examples each, as shown in the figure. Using $\mathrm{CI}(B_1) = \mathrm{CI}(B_2) = 1$, one can easily check that the output of Algorithm 1 will be $\mathcal{B} = \{(B_1, 1), (B_2, 1)\}$.

Now consider the 1-LC block system $\mathcal{B}' = \{(B', 1)\}$, where B' consists of the 10 **red** and 3 **purple** points as shown in Figure 1. It is an 1-LC block system, E^+-relevant, consistent, and $\mathcal{B} \succ_{E^+} \mathcal{B}'$. Hence, $\mathcal{B}$ is not block-minimal consistent. □

Note that B' in the example is not convex. In the theorem below, we give a sufficient condition for Algorithm 1 to return block-minimal consistent hypohteses. It requires, among others, that the underlying metric space is blockwise convex. However, whether this condition is also necessary remains an open question.

Theorem 5. *Let $\mathcal{M} = (X, D)$ be a blockwise convex metric space for some $\tau \geq 0$. If there exists some constant $\sigma > 0$ such that $\mathrm{CI}(B) = \sigma$ for all θ'-convex blocks $B \subseteq X$ with $\theta' \geq \tau$, then the output of Algorithm 1 is block-minimal consistent.*

5 Application: Learning Weakly Convex Boolean Functions

To demonstrate the performance of our general-purpose heuristic LCM in practice, we consider the CHF problem for the special case where the underlying metric space is the Hamming space $\mathcal{M}_H = (\mathbf{B}_d, D_H)$. This case gives rise to weakly convex Boolean functions, i.e., whose sets of true points are θ-convex [14]. As previously discussed, such functions can be represented by DNFs, where the terms correspond to the θ-convex blocks of their sets of true points. The CHF problem for this class of Boolean functions was studied and solved in [14], using a *domain-specific* algorithm that computes consistent θ-GC hypotheses.[4] One can verify that $\mathcal{M}_H$ is *blockwise convex* for all $\theta \geq 2$ (cf. [13]) and that $\mathrm{CI}(B) = 1$ for all $\theta \geq 2$ and all θ-convex convex blocks B, leading to the following result by Theorem 5:

Corollary 1. *For $\mathcal{M} = \mathcal{M}_H$ and $\tau = 2$, all θ-LC block systems returned by Algorithm 1 are block-minimal consistent.*

For all experiments, we set $\tau = 2$ based on the rationale discussed above. The application of Algorithm 1 to weakly convex Boolean functions involves the implementation of the following subroutines for (irredundant) terms B and R over variables $x_1, \ldots, x_d$ ($m_\oplus$ and $m_\ominus$ below denote $|E^+|$ and $|E^-|$ in Algorithm 1, respectively):

- WEAKLYCONVEXHULL(θ, E^+) computes a set of conjunctive terms representing the blocks of $\rho_\theta(E^+)$ in $\mathcal{O}(dm_\oplus^2)$ time (see [14] for details).
- MEMBERSHIP(e, B) determines in $\mathcal{O}(d)$ time whether $e \in \mathbf{B}_d$ satisfies B.
- DISTANCE(B, R) computes the distance between the two subcubes of $\mathbf{B}_d$ represented by B and R in $\mathcal{O}(d)$ time.
- CONNECTIVITYINDEX(B) returns 1 in constant time (see the remark above).
- JOIN(θ, B, R) computes the conjunction representing the *smallest* subcube of $\mathbf{B}_d$ containing the subcubes represented by B and R in $\mathcal{O}(d)$ time.

Since all functions run in time polynomial in $m_\oplus$ and the parameter d, Algorithm 1 runs in time polynomial in $m_\oplus$, $m_\ominus$, and d by Theorem 1 (ii).

5.1 Experimental Results

In this section, we present our experimental results on learning θ-convex Boolean functions. We empirically compare the *number of blocks* and the *predictive accuracy* of the output of Algorithm 1 with those of two baseline methods. The first one is the algorithm

[4] In [14], the authors use the notation k instead of θ and refer to k-convex Boolean functions.

in [14], developed specifically for this task. It computes a DNF representing the consistent θ-GC hypothesis. As the second baseline method, we compare the results with the DNFs extracted from Boolean decision trees learned on the same training data. The rationale for considering this baseline is that Boolean decision trees represent DNFs.

Datasets For each DNF learning task, we first generated t conjunctions, each with the following procedure for parameters $\tilde{d}, \theta_{\text{min}}, \theta_{\text{max}} \in \mathbb{N}$: Tossing a biased coin $\tilde{d}$ times independently and with success probability $\ell/\tilde{d}$ for some random integer $0 \leq \ell \leq \tilde{d}$, we have generated a subset V of the Boolean variables $x_1, \ldots, x_{\tilde{d}}$. For each $x_i \in V$, we then tossed an unbiased coin and, depending on the outcome, added either x_i or $\overline{x}_i$ to the conjunction. For each pair of the t conjunctions, a distance $\theta' \in \mathbb{N}$ is generated uniformly at random within the interval $[\theta_{\text{min}}, \theta_{\text{max}}]$. It determines the number of conflicting new variables–distinct from those in V–that must be added to both terms. Thus, after processing all pairs, we obtain a θ-convex DNF over $d \geq \tilde{d}$ Boolean variables for some $\theta \geq \theta_{\text{min}}$, consisting of t terms. This DNF was used as the *unknown* target weakly convex Boolean function. Finally, the positive (resp. negative) training examples E^+ (resp. E^-) of varying sizes were chosen *uniformly* from this DNF's true (resp. false) points.

Parameters We considered all combinations of $t \in \{4, 5\}$, $\tilde{d} \in \{10, 15, 20\}$, and $\ell \in \{4, 5\}$. $\theta_{\text{min}} = 3$ and $\theta_{\text{max}} = 6$ were constant as a compromise between variability of the distances and dimensionality of the underlying Hamming space. Regarding the training examples, we considered two cases, the *balanced* with $|E^+| = |E^-|$ and the *imbalanced* one with $|E^-| = 150000 \gg |E^+|$. In both cases, $|E^+| \in \{10, 20, \ldots, 100, 200, \ldots, 1000, 1500, 2000, \ldots, 5000\}$. In order to estimate mean and standard deviation of the performance measures, the experiment was repeated $i = 50$ times, independently for each parameter combination.

Limitations The experimental design described above has some inherent *limitations*. Most importantly, due to the addition of conflicting variables during the concept generation, the *dimension* d of the underlying Hamming spaces is also determined randomly. However, as the VC-dimension of weakly convex Boolean functions is tied to the dimension of the surrounding space, this has a direct impact on learnability. In particular, none of the parameter combinations we considered satisfies any bounds for efficient PAC-learnability [14]. This limits the number of terms that can be considered, as in this case *much* more training examples are needed. Interestingly, the number of *negative* examples appears to govern this effect, which was investigated in the *imbalanced* case described above. Furthermore, it is known that there is an inherent imbalance between true and negative points in weakly convex Boolean functions [14]. This is reflected neither in the balanced nor in the imbalanced cases described above, as otherwise we would end up with high probability with none or only very few positive examples. Notice that the DNF generation is biased also in the sense that the conflicting variables are disjoint for distinct term pairs, except for the initial $\tilde{d}$ common variables. Increasing the overlap of common conflicting variables results in smaller blocks, further emphasizing the imbalance between true and negative points of the generated DNFs. As mentioned, the goal of our experiments is to examine, as a proof-of-concept, Algorithm 1 in terms of *compactness* (number of blocks) and *predictive* performance by comparing its output hypotheses to those of the two baseline algorithms. Accordingly, since our general-

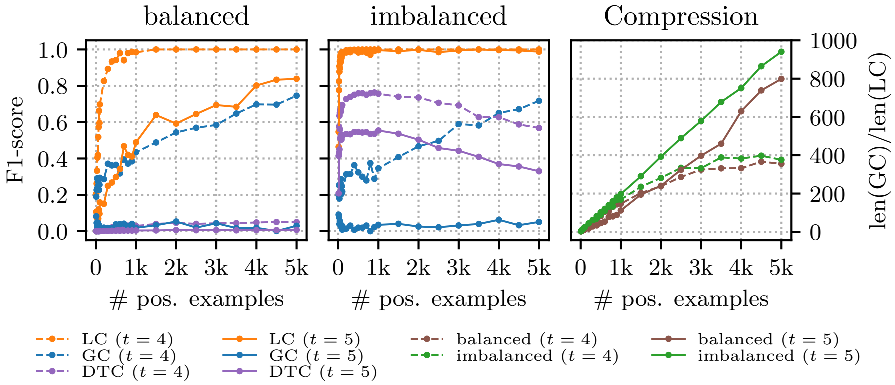

Fig. 2. Mean F1-score obtained by θ-LC (**orange** lines), θ-GC (**blue** lines), and DTC hypotheses (**purple** lines) depending on the number $|E^+|$ of positive examples drawn from target concepts with $t = 4$ (dashed lines) and $t = 5$ terms (solid lines). The *balanced* case (left plot) with $|E^-| = |E^+|$ is distinguished from the *imbalanced* case (center plot) where $|E^-| = 150000 \gg |E^+|$ is constant. The compactness (right plot) is the ratio of the θ-GC over θ-LC hypotheses' number of blocks (i.e., terms in the output DNFs). All values are averaged over 50 independent iterations.

purpose algorithm does not utilize any domain-specific knowledge, except for setting $\tau = 2$, a comparison with state-of-the-art algorithms specific to learning DNFs is out of the scope of this work. For all learning tasks and iterations, the decision tree model's hyperparameters were individually optimized using cross-validated grid search.

Results For each learning task, the three algorithms are called with E^+ and E^-, using $\tau = 2$ for the first two algorithms. The corresponding three DNFs are denoted by ϕ_{GC} (consistent θ-GC hypothesis, i.e., the largest consistent θ-convex hull of E^+ [14]), ϕ_{LC} (consistent θ-LC block system produced by Algorithm 1), and ϕ_{DTC} (decision tree). The hypotheses are compared with each other by their number of terms (i.e., blocks) and F1-score. The results[5] for $t \in \{4, 5\}$ and $\tilde{d} = 15$ are shown in Figure 2. Notice that the output hypotheses of all three algorithms show higher predictive performance for target concepts with $t = 4$ terms (dashed lines) compared to $t = 5$ (solid lines). This is expected because a) there are more examples *per block* for $t = 4$ than for $t = 5$, and b) the dimension d of the Hamming space increases with t due to the addition of conflicting variables to the terms. Specifically, d ranged from 27 to 48 (mean 37.84, std. dev. 3.82) for $t = 4$ and from 36 to 66 (mean 52.92, std. dev. 5.29) for $t = 5$.

Notice that the DNFs extracted from decision tree classifiers (DTC) for the balanced case (left plot) performed very poorly, regardless of t and the number of examples. However, $|E^+| = |E^-| \geq 1000$ examples suffice for Algorithm 1 to return excellent

[5] The algorithms and the experiments were implemented in Python 3.11 using the `sortedcontainers` package for managing the underlying data structures of Algorithm 1 and the θ-convex baseline algorithm [23]. For the decision tree models, we used the implementation of the `sklearn` package.

θ-LC hypotheses with a stable F1-score of almost 1.0 (mean 0.998, std. dev. 0.037) for target concepts with $t = 4$ terms (dashed lines). For $t = 4$ and $|E^+| \geq 1000$, the returned hypotheses even coincide *exactly* with the unknown target concepts in more than 95.4% of the cases; for $t = 5$ and $|E^+| \geq 1000$, in about 52.89% of the cases. The average F1-score also drops substantially in the latter case ($t = 5$) but was still 0.84 (std. dev. 0.32) for $|E^+| = |E^-| = 5000$ positive and negative examples. In contrast, the θ-GC hypotheses performed far worse than θ-LC hypotheses. In particular, they coincided with the unknown target concepts exactly only in 19.6% of the cases with an average F1-score of 0.75 (std. dev. 0.39) for $t = 4$ and not even once with an average F1-score of only 0.03 (std. dev. 0.15) for $t = 5$, for $|E^+| \geq 1000$ in both cases. It is worth mentioning that the θ-GC hypotheses have a very high precision (near 1.0 almost all the time) but a poor mean recall of 0.03 (std. dev. 0.13). This is a direct consequence of the effect of the global distance constraint θ, which prevents the necessary join operations leading to hypotheses with several very small blocks, often even only singletons. In other words, the θ-GC hypotheses do *not* generalize at all from the training data.

A comparison between the balanced (left plot) and the imbalanced case (center plot) reveals that the predictive performance of θ-GC hypotheses is not affected by the additional negative examples. This is to be expected because, as discussed before, θ-GC hypotheses appear to often overfit E^+. In contrast, since Algorithm 1 greedily joins blocks until inconsistency with E^-, it benefits more from the additional negative examples. It is remarkable, that it obtains an excellent F1-score of 0.99 (std. dev. 0.09) even for $t = 5$ terms when provided with $|E^+| \approx 1000$ positive and $|E^-| = 150000$ negative examples. Another difference to the balanced case is that the DNFs extracted from decision tree classifiers also appear to benefit from the additional negative examples. Still, they perform significantly worse than the θ-LC hypotheses for both $t = 4$ (mean 0.68, std. dev. 0.22) and $t = 5$ (mean 0.48, std. dev. 0.32).

The right plot in Figure 2 shows the mean ratio of the lengths (i.e., number of blocks) of θ-GC hypotheses over θ-LC hypotheses. θ-GC hypotheses have up to almost *three* orders of magnitudes more terms than θ-LC hypotheses. On average, the factor is 141.13 (std. dev. 245.38) for unknown target concepts with $t = 4$ and 230.04 (std. dev. 293.29) with $t = 5$ terms.

In summary, our experiments show that Algorithm 1 solves the CHF problem for weakly convex DNFs with *significantly less* blocks and with a (much) *better* average predictive performance compared to the related baseline decision tree and weakly convex DNF learning algorithms [14, 23].

6 Concluding Remarks

Weak convexity [14, 23] has proven to be a powerful parameterized tool for solving the CHF problem for hypotheses composed of pairwise separated blocks. A major limitation of the approaches in [14, 23] is that the pairwise distances between blocks of consistent θ-GC hypotheses are often determined by the local configuration of only a few training examples. As our experimental results in Section 5.1 demonstrate, this can lead to *poor* generalization performance. To address this issue, we introduced and studied LC block systems, a general framework for discontiguous hypothesis classes that extends weakly convex hulls in finite metric spaces. Motivated by the negative complexity result in Theorem 3, we proposed a greedy heuristic to compute consistent and compact LC block systems.

For simplicity, this short version restricts the discussion to geodesic convexity, a special case of *interval convexity* [4]. In addition to the Hamming space considered in this paper, this special case covers other metric spaces commonly used in machine learning, such as those formed by the vertex set of a graph equipped with the shortest-path or weighted shortest-path distance (see, e.g., [3, 12, 22, 24]).

In the special case of learning weakly convex Boolean functions, our heuristic LCM is optimal, meaning that no coarser consistent LC block system exists with fewer blocks than the output of our algorithm. Our experimental results clearly show that the hypotheses generated by our *general-purpose* heuristic achieve *significantly better* predictive performance compared to those produced by the domain-specific method in [14] and by Boolean decision tree learning algorithms. The improvement over [14] can largely be attributed to the *compactness* of the output hypotheses: our approach generates hypotheses with significantly fewer blocks than those produced by the method in [14]. In fact, they are *near-optimal* in terms of the number of blocks in most cases.

The approach and results of this paper raise several questions for further research. For instance, is the sufficient condition of block-minimal consistency in Theorem 5 also necessary? If not, what properties characterize this kind of optimality? Another interesting question is whether our heuristic can be adapted to *unsupervised* learning problems. This question is motivated by the strong relationship between LC block systems and *density-based clusters* [5, 15], which share similar definitions of connectedness, global parameters that limit expressivity, and similar algorithmic strategies for greedily joining blocks or clusters by ascending distance.

Another promising avenue for future research could involve *relaxing* the strictness of LC block systems to tolerate a certain amount of misclassifications, akin to soft margin support vector machines [10]. Additionally, motivated by various learning problems over infinite domains, extending the results of this paper from finite to *infinite* metric spaces presents an important and challenging task. This extension is nontrivial, requiring a careful integration of concepts from topology, computational complexity, and machine learning.

Acknowledgements We are grateful to Victor Chepoi for insightful discussions on abstract convexity.

References

1. Bereg, S., Cabello, S., Díaz-Báñez, J.M., Pérez-Lantero, P., Seara, C., Ventura, I.: The class cover problem with boxes. Computational Geometry **45**(7), 294–304 (2012)
2. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM **36**(4), 929–965 (1989)
3. Bressan, M., Cesa-Bianchi, N., Lattanzi, S., Paudice, A.: Exact recovery of clusters in finite metric spaces using oracle queries. In: Belkin, M., Kpotufe, S. (eds.) Proc. of Thirty Fourth Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 134, pp. 775–803. PMLR, Boulder, Colorado, USA (2021)

4. Calder, J.R.: Some elementary properties of interval convexities. Journal of the London Mathematical Society **s2-3**(3), 422–428 (1971)
5. Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J.: Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data **10**(1), 5:1–5:51 (2015)
6. Chalopin, J., Chepoi, V., Moran, S., Warmuth, M.K.: Unlabeled sample compression schemes and corner peelings for ample and maximum classes. Journal of Computer and System Sciences **127**, 1–28 (2022)
7. Chepoi, V.: Classification of graphs by means of metric triangles. Metody Diskretnogo Analiza **96**, 75–93 (1989)
8. Chepoi, V.: Basis graphs of even delta-matroids. Journal of Combinatorial Theory, Series B **97**(2), 175–192 (2007)
9. Chepoi, V., Knauer, K., Marc, T.: Hypercellular graphs: Partial cubes without Q3- as partial cube minor. Discrete Mathematics **343**(4), 111678 (2020)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
11. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, Cambridge, UK, 2nd edn. (2002)
12. de Araújo, P.H.M., Campêlo, M.B., Corrêa, R.C., Labbé, M.: The geodesic classification problem on graphs. Electronic Notes in Theoretical Computer Science **346**, 65–76 (2019)
13. Ekin, O., Hammer, P.L., Kogan, A.: On connected Boolean functions. Discrete Applied Mathematics **96-97**, 337–362 (1999)
14. Ekin, O., Hammer, P.L., Kogan, A.: Convexity and logical analysis of data. Theoretical Computer Science **244**(1), 95 – 116 (2000)
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. AAAI Press (1996)
16. Hill, R.: A First Course in Coding Theory. Oxford Applied Mathematics and Computing Science Series, Clarendon Press, Oxford (1986)
17. Horváth, T., Turán, G.: Learning logic programs with structured background knowledge. Artificial Intelligence **128**(1-2), 31–97 (2001)
18. Kietz, J., Dzeroski, S.: Inductive logic programming and learnability. SIGART Bull. **5**(1), 22–32 (1994)
19. Lantero, P.P.: Geometric Optimization for Classification Problems. PhD Theis, Universidad de Sevilla (2010)
20. Mitchell, T.M.: Generalization as search. Artificial intelligence **18**(2), 203–226 (1982)
21. Natarajan, B.K.: On learning Boolean functions. In: Proc. of the Nineteenth Annual ACM Symposium on Theory of Computing. pp. 296–304. STOC '87, Association for Computing Machinery, New York City, NY, USA (1987)
22. Seiffarth, F., Horváth, T., Wrobel, S.: Maximal closed set and half-space separations in finite closure systems. Theoretical Computer Science **973**, 114105 (2023)
23. Stadtländer, E., Horváth, T., Wrobel, S.: Learning weakly convex sets in metric spaces. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano, J.A. (eds.) Machine Learning and Knowledge Discovery in Databases. Research Track. LNCS, vol. 12976, pp. 200–216. Springer, Cham (2021)
24. Thiessen, M., Gärtner, T.: Online learning of convex sets on graphs. In: Amini, M.R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 349–364. Springer, Cham (2022)
25. van de Vel, M.L.J.: Theory of Convex Structures, North-Holland Mathematical Library, vol. 50. North-Holland, Amsterdam (1993)