A Comparison of Lex Bounds for Multiset Variables in Constraint Programming

Y.C. Law[1] and **J.H.M. Lee**[1] and **T. Walsh**[2] and **M.H.C. Woo**[1]
[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
{yclaw,jlee,hcwoo}@cse.cuhk.edu.hk
[2]NICTA & UNSW
Sydney, Australia
toby.walsh@nicta.com.au

Abstract

Set and multiset variables in constraint programming have typically been represented using subset bounds. However, this is a weak representation that neglects potentially useful information about a set such as its cardinality. For set variables, the length-lex (LL) representation successfully provides information about the length (cardinality) and position in the lexicographic ordering. For multiset variables, where elements can be repeated, we consider richer representations that take into account additional information. We study eight different representations in which we maintain bounds according to one of the eight different orderings: length-(co)lex (LL/LC), variety-(co)lex (VL/VC), length-variety-(co)lex (LVL/LVC), and variety-length-(co)lex (VLL/VLC) orderings. These representations integrate together information about the cardinality, variety (number of distinct elements in the multiset), and position in some total ordering. Theoretical and empirical comparisons of expressiveness and compactness of the eight representations suggest that length-variety-(co)lex (LVL/LVC) and variety-length-(co)lex (VLL/VLC) usually give tighter bounds after constraint propagation. We implement the eight representations and evaluate them against the subset bounds representation with cardinality and variety reasoning. Results demonstrate that they offer significantly better pruning and runtime.

Introduction

In constraint programming, we often need to model multisets (or bags) of objects. For example, in the template design problem (prob002 in CSPLib (Gent and Walsh 1999)), we need to construct printing templates, which are multisets of different designs. Multisets, unlike sets, can contain repetition of elements. For popular designs, we may have multiple copies on the same template. Surprisingly, whilst there has been significant progress on developing representations for sets, relatively little research has been done on how best to represent multisets.

Sadler and Gervet (2004) proposed representing set variables with subset, lexicographic, and cardinality bounds. Indeed, they suggested that such a representation could also be used for multisets (2008). However, little detail is provided about how to do this exactly. To compare two multisets, they lexicographically compare their occurrence vectors written in decreasing order. For instance, $\{3,3,2,1,1\} \preceq \{4\} \preceq \{4,4\}$. Gervet and Van Hentenryck (2006) proposed representing set variables using length-lex bounds, arguing that it provides comparable pruning to the aforementioned hybrid domains at a fraction of the computational cost. It is therefore promising to consider length-lex and related bounds for multiset variables. However, as a number of different orderings are possible, we have undertaken a theoretical and empirical comparison of the most promising options.

As multisets permit repeated elements, we can incorporate information about the variety (number of distinct elements) (Law, Lee, and Woo 2009) in addition to the cardinality and position in the lexicographic ordering. As a result, we introduce eight different representations for multiset variables in which we maintain bounds according to one of eight different orderings: length-(co)lex (LL/LC), variety-(co)lex (VL/VC), length-variety-(co)lex (LVL/LVC), and variety-length-(co)lex (VLL/VLC) orderings. These bounds provide information about the possible cardinality, variety, and position in the (co)lexicographic ordering of a multiset. We evaluate the expressiveness (whether the set of multisets can be exactly represented) and compactness (whether the interval is minimal) of the eight representations both theoretically and empirically. Our results suggest that LVL/LVC and VLL/VLC representations are usually more expressive and more compact than LL/LC and VL/VC respectively. The eight representations give total orderings on multisets, which make enforcing bounds consistency on multiset variables possible. However, when we attempt to enforce bounds consistency on the bounds of the proposed representations, this operation can be NP-hard even on unary constraints. To test out these representations, we implement the eight representations and evaluate them against the subset bounds representation with cardinality and variety reasoning. Results confirm that these new representations achieve significantly better pruning and runtime.

Background

Set Variables

A *set* is an unordered list of elements *without repetition*. The *cardinality* of a set S is the number of elements in S, denoted as $|S|$. Gervet (1997) proposed to represent the domain of a set variable S with an interval $[glb(S), lub(S)]$

such that $D_S = \{m \mid glb(S) \subseteq m \subseteq lub(S)\}$. The *greatest lower bound* $glb(S)$ contains all the elements which *must exist* in the set, while the *least upper bound* $lub(S)$ contains any element which *can exist* in the set. S is said to be *bound* when its lower bound equals its upper bound (i.e., $glb(S) = lub(S)$). In this subset bounds representation, the set domain is ordered partially under $\subseteq$. It also neglects the cardinality and the position in lexicographic ordering which can be important in many problems. Thus, Gervet and Van Hentenryck (2006) proposed to *totally* order a set domain with a length-lex ordering. This representation incorporates the cardinality and the position in lexicographic ordering directly, giving tighter bounds when enforcing bounds consistency.

Notation Given a universe U of integers $\{1, \ldots, n\}$, set variables, denoted as S_i, takes their values from U. Sets are denoted by letters s, t, x, and y. A subset s of U of cardinality c is denoted by $\{s_1, s_2, \ldots, s_c\}$ where $s_1 < s_2 < \cdots < s_c$.

Length-lex Ordering The length-lex ordering $\preceq$ totally orders sets first by cardinality and then lexicographically.

Definition 1. *A length-lex ordering $\preceq$ is defined by:* $s \preceq t$ *iff* $s = \emptyset \vee |s| < |t| \vee (|s| = |t| \wedge (s_1 < t_1 \vee s_1 = t_1 \wedge s \setminus \{s_1\} \preceq t \setminus \{t_1\}))$.

Definition 2. *Given a universe U, a length-lex interval is a pair of sets $\langle m, M\rangle$ which represents the sets between m and M in the length-lex ordering (i.e., $\{s \subseteq U \mid m \preceq s \preceq M\}$).*

Given a universe $U = \{1, \ldots, 4\}$, the sets are ordered as follows: $\emptyset \preceq \{1\} \preceq \{2\} \preceq \{3\} \preceq \{4\} \preceq \{1,2\} \preceq \{1,3\} \preceq \{1,4\} \preceq \{2,3\} \preceq \{2,4\} \preceq \{3,4\} \preceq \{1,2,3\} \preceq \{1,2,4\} \preceq \{1,3,4\} \preceq \{2,3,4\} \preceq \{1,2,3,4\}$. The length-lex interval $\langle\{1,2\},\{3,4\}\rangle$ denotes the set $\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\}$.

Multiset Variables

A *multiset* is a generalization of set that allows elements to repeat. Without loss of generality, we assume that multiset elements are positive integers from 1 to n. We shall use $\emptyset$ to denote both the empty set and the empty multiset. The universe of a multiset is a multiset itself, which defines the maximum possible occurrences of each element. Given a universe U, we denote a multiset S as $S = \{\!\{m_1, m_2, \cdots, m_c\}\!\}$ where $m_i \leq m_j$ for $1 \leq i \leq j \leq c$, its cardinality (total number of elements) as $|S|$, and its variety (total number of distinct elements) (Law, Lee, and Woo 2009) as $\|S\|$. For example, if $S = \{\!\{1,1,2,2,3\}\!\}$, then $|S| = 5$ and $\|S\| = 3$. Since an element in a multiset variable can occur multiple times, we let $occ(i, S)$ be the number of occurrences of an element i in the multiset S. Walsh (2003) proposed using an occurrence vector $\langle occ(1,S), \ldots, occ(n,S)\rangle$ to represent a multiset variable with n elements. For example, the occurrence representation for the value $\{\!\{1,1,2,2,3\}\!\}$ with the universe $U = \{\!\{1,1,2,2,3,3\}\!\}$ is $\langle occ(1,S), occ(2,S), occ(3,S)\rangle = \langle 2,2,1\rangle$.

Note that a set value can also be represented using the occurrence representation in which the number of occurrence is either 0 or 1 to denote the existence of the corresponding element. Thus, we adopt the occurrence representation for multiset variables and order the occurrence vector to give various orderings in multisets.

Lex-induced Orderings in Multisets

The length-lex representation for sets incorporates information about the length (cardinality) and position in the lexicographic ordering. Such a representation can be extended to include the variety information since multisets allow repeated elements. This gives a total of eight different ways to order multisets. In the following, we formally define the eight orderings, in which four of them order the position lexicographically and the other four colexicographically.

Lex Orderings

The *lex ordering* $\preceq_l$ totally orders multisets *lexicographically*. Here, we assume the multisets are represented by the occurrence representation (i.e., the number of occurrences of each element are stored in an occurrence vector). Thus, given two multisets x and y, we compare their occurrence vectors $\langle occ(1,x), \ldots, occ(n,x)\rangle$ and $\langle occ(1,y), \ldots, occ(n,y)\rangle$ from the *first* position to the *last*.

Definition 3. *A* lex ordering $\preceq_l$ *is defined by:* $x \preceq_l y$ *iff* $(x = y) \vee (\exists i, occ(i,x) < occ(i,y) \wedge \forall j < i, occ(j,x) = occ(j,y))$.

For example, consider two multisets $x = \{\!\{1,2,2\}\!\}$ and $y = \{\!\{1,3,3\}\!\}$. Their occurrence vectors are $\langle 1,2,0\rangle$ and $\langle 1,0,2\rangle$ respectively. $\{\!\{1,3,3\}\!\} \preceq_l \{\!\{1,2,2\}\!\}$ because $occ(1,y) = occ(1,x)$ and $occ(2,y) < occ(2,x)$.

Colex Orderings

Contrary to the lex ordering, the *colex ordering* $\preceq_c$ compares the occurrence vectors of two multisets from the *last* position to the *first*.

Definition 4. *A* colex ordering $\preceq_c$ *is defined by:* $x \preceq_c y$ *iff* $(x = y) \vee (\exists i, occ(i,x) < occ(i,y) \wedge \forall j > i, occ(j,x) = occ(j,y))$.

For example, let two multisets $x = \{\!\{1,3,3\}\!\}$ and $y = \{\!\{2,3,3\}\!\}$ with occurrence vectors $\langle 1,0,2\rangle$ and $\langle 0,1,2\rangle$ respectively. They are ordered as $\{\!\{1,3,3\}\!\} \preceq_c \{\!\{2,3,3\}\!\}$ because $occ(3,x) = occ(3,y)$ and $occ(2,x) < occ(2,y)$.

Induced Orderings

Given a total order $\preceq_\beta$ on a set of multisets, we can have four different $\preceq_\beta$-induced orderings when we integrate $\preceq_\beta$ with cardinality and/or variety of multisets.

Length-β Ordering The *length-β ordering* $\preceq_{l\beta}$ totally orders multisets first by their cardinality, and then by the β ordering: $x \preceq_{l\beta} y$ iff $|x| < |y| \vee (|x| = |y| \wedge x \preceq_\beta y)$.

Variety-β Ordering The *variety-β ordering* $\preceq_{v\beta}$ totally orders multisets first by their variety, and then by the β ordering: $x \preceq_{v\beta} y$ iff $\|x\| < \|y\| \vee (\|x\| = \|y\| \wedge x \preceq_\beta y)$.

Length-β and variety-β prefer cardinality and variety over the β ordering respectively. In fact, both cardinality and variety can be considered together, giving two more orderings.

Length-variety-β Ordering The *length-variety-β ordering* $\preceq_{lv\beta}$ totally orders multisets first by their cardinality, then by their variety, and then by the β ordering: $x \preceq_{lv\beta} y$ iff $|x| < |y| \vee (|x| = |y| \wedge x \preceq_{v\beta} y)$.

Variety-length-β Ordering The *variety-length-β ordering* $\preceq_{vl\beta}$ totally orders multisets first by their variety, then by their cardinality, and then by the β ordering: $x \preceq_{vl\beta} y$ iff $\|x\| < \|y\| \vee (\|x\| = \|y\| \wedge x \preceq_{l\beta} y)$.

Since lex and colex orderings are total orders, we can have eight different orderings by substituting β by the lex and colex orderings. For example, substituting β by the lex ordering in the length-β ordering gives the length-lex ordering LL ($\preceq_{ll}$). Similarly, we can have variety-lex VL ($\preceq_{vl}$), length-variety-lex LVL ($\preceq_{lvl}$), variety-length-lex VLL ($\preceq_{vll}$), length-colex LC ($\preceq_{lc}$), variety-colex VC ($\preceq_{vc}$), length-variety-colex LVC ($\preceq_{lvc}$), and variety-length-colex VLC ($\preceq_{vlc}$) orderings.

The above eight orderings are applicable to multisets. All the four colex orderings on multisets reduce to the LL ordering on sets introduced by Gervet and Van Hentenryck (2006). Note that, when we consider a fixed length, the colex (*resp.* lex) ordering for set values is equivalent to ordering the occurrence vector lexicographically (*resp.* colexicographically).

The domain of a multiset variable is simply a set of multisets. We can thus totally order the domain values of a variable according to the eight orderings. To illustrate the differences, Table 1 lists the domain of a multiset variable S with universe $U = \{\!\{1, 2, 2, 3, 3\}\!\}$ in the four lex orderings. Take the LVL ordering as an example. We first order the multisets by their cardinality. Thus, $\emptyset$ has cardinality 0 and is the first multiset, followed by the multisets with cardinalities 1, 2, and so on. For multisets of the same cardinality, we then compare their variety. Consider the segment with cardinality 2, i.e., from $\{\!\{3, 3\}\!\}$ to $\{\!\{1, 2\}\!\}$. The multisets $\{\!\{3, 3\}\!\}$ and $\{\!\{2, 2\}\!\}$ are ordered before $\{\!\{2, 3\}\!\}$, $\{\!\{1, 3\}\!\}$, and $\{\!\{1, 2\}\!\}$ because the former two have variety 1 and the latter ones have variety 2. Lastly, we order the multisets lexicographically. The occurrence vectors of $\{\!\{3, 3\}\!\}$ and $\{\!\{2, 2\}\!\}$ are $\langle 0, 0, 2\rangle$ and $\langle 0, 2, 0\rangle$ respectively. Thus, $\{\!\{3, 3\}\!\} \preceq_{lvl} \{\!\{2, 2\}\!\}$ because $occ(1, \{\!\{3, 3\}\!\}) = occ(1, \{\!\{2, 2\}\!\}) = 0$ and $occ(2, \{\!\{3, 3\}\!\}) < occ(2, \{\!\{2, 2\}\!\})$.

Given a multiset variable, we can approximate its domain, which is a set S of multisets, with an *α-interval*, where α refers to one of the above eight orderings. The interval $\langle m, M\rangle_\alpha$ must contain all the multisets in S such that m and M are the lower and upper bounds of S respectively. We also define the *α-closure* of S which is the minimal possible α-interval containing S.

Definition 5. *Given an α ordering, an* α-interval $\langle m, M\rangle_\alpha$ *is a set of multisets defined by $\langle m, M\rangle_\alpha = \{x \mid m \preceq_\alpha x \preceq_\alpha M\}$. The* α-closure *of S is defined by $cl_\alpha(S) = \langle m, M\rangle_\alpha$, where $S \subseteq \langle m, M\rangle_\alpha$ and there does not exist $m \prec_\alpha m'$ and $M' \prec_\alpha M$ such that ($m \neq m'$ or $M \neq M'$) and $S \subseteq \langle m', M'\rangle_\alpha$.*

Definition 6. *An α* representation *of a set S of multisets is $cl_\alpha(S)$. An α representation of S is* exact *if $S = cl_\alpha(S)$.*

For example, let the universe $U = \{\!\{1, 2, 2, 3, 3\}\!\}$ and $S = \{\{\!\{1\}\!\}, \{\!\{2, 2\}\!\}, \{\!\{2, 3\}\!\}\}$. The lvl-closure of S is the lvl-interval $\langle\{\!\{1\}\!\}, \{\!\{2, 3\}\!\}\rangle_{lvl}$. This representation is *not* exact, as the interval contains the multiset $\{\!\{3, 3\}\!\} \notin S$.

Expressiveness

An exact representation gives the tightest possible bounds and contains no undesired values. It is often the case that a set of multisets can be exactly represented using one representation but not using a different representation. In this section, we compare the eight representations to see which ordering is better in terms of the notion "expressiveness".

Definition 7. *(Walsh 2003) Given a universe U and two different multiset representations A and B. A is said to be* as expressive as *B if $\forall S \subseteq U, (S = cl_A(S)) \leftrightarrow (S = cl_B(S))$. A is said to be* more expressive *than B if $\forall S \subseteq U, (S = cl_B(S)) \rightarrow (S = cl_A(S))$ and $\exists S \subseteq U, (S = cl_A(S)) \wedge (S \neq cl_B(S))$. A and B are* incomparable *if neither one of them is more expressive than the other.*

The following propositions compare the expressiveness of the eight representations under the conditions that the cardinality and/or variety of a set of multisets is fixed.

Proposition 1. *When both the cardinality and variety are fixed, (i) the LVL/LVC representation is as expressive as the VLL/VLC representation, (ii) the LVL/LVC and VLL/VLC representations are more expressive than the LL/LC and VL/VC representations respectively, and (iii) the LVL is as expressive as the LVC and the VLL is as expressive as the VLC.*

The results in Proposition 1 can be demonstrated using the example in Table 1. When the cardinality and variety are 2 and 1 respectively, the LVL and VLL representations can exactly represent $\{\{\!\{2, 2\}\!\}, \{\!\{3, 3\}\!\}\}$ by the lvl-interval $\langle\{\!\{3, 3\}\!\}, \{\!\{2, 2\}\!\}\rangle_{lvl}$ and the vll-interval $\langle\{\!\{3, 3\}\!\}, \{\!\{2, 2\}\!\}\rangle_{vll}$ respectively. However, the LL and VL representations give the ll-interval $\langle\{\!\{3, 3\}\!\}, \{\!\{2, 2\}\!\}\rangle_{ll}$ and the vl-interval $\langle\{\!\{3, 3\}\!\}, \{\!\{2, 2\}\!\}\rangle_{vl}$ respectively, in which both contain the additional undesired value $\{\!\{2, 3\}\!\}$.

The following two propositions relax the conditions to the case that either the cardinality or the variety is fixed.

Proposition 2. *When the cardinality is fixed, (i) the LVL/LVC representation is more expressive than the VLL/VLC, LL/LC, and VL/VC representations, and (ii) the LL representation is as expressive as the LC representation.*

Proposition 3. *When the variety is fixed, (i) the VLL/VLC representation is more expressive than the LVL/LVC, LL/LC, and VL/VC representations, and (ii) the VL representation is as expressive as the VC representation.*

In Table 1, when the cardinality is 3, the LVL representation can exactly represent the multisets by the lvl-interval $\langle\{\!\{2, 3, 3\}\!\}, \{\!\{1, 2, 3\}\!\}\rangle_{lvl}$, while the VLL, LL, or VL representations cannot. There are additional undesired values in their corresponding intervals. In fact, when only the variety is fixed, we obtain similar results. Suppose the variety is 2, the VLL representation can exactly represent the multisets

Table 1: The four lex orderings for the domain of a multiset variable S with universe $U = \{\!\{1, 2, 2, 3, 3\}\!\}$

Length-lex (LL)	$\emptyset \preceq_{ll} \{\!\{3\}\!\} \preceq_{ll} \{\!\{2\}\!\} \preceq_{ll} \{\!\{1\}\!\} \preceq_{ll} \{\!\{3,3\}\!\} \preceq_{ll} \{\!\{2,3\}\!\} \preceq_{ll} \{\!\{2,2\}\!\} \preceq_{ll} \{\!\{1,3\}\!\}$ $\preceq_{ll} \{\!\{1,2\}\!\} \preceq_{ll} \{\!\{2,3,3\}\!\} \preceq_{ll} \{\!\{2,2,3\}\!\} \preceq_{ll} \{\!\{1,3,3\}\!\} \preceq_{ll} \{\!\{1,2,3\}\!\} \preceq_{ll} \{\!\{1,2,2\}\!\}$ $\preceq_{ll} \{\!\{2,2,3,3\}\!\} \preceq_{ll} \{\!\{1,2,3,3\}\!\} \preceq_{ll} \{\!\{1,2,2,3\}\!\} \preceq_{ll} \{\!\{1,2,2,3,3\}\!\}$
Variety-lex (VL)	$\emptyset \preceq_{vl} \{\!\{3\}\!\} \preceq_{vl} \{\!\{3,3\}\!\} \preceq_{vl} \{\!\{2\}\!\} \preceq_{vl} \{\!\{2,2\}\!\} \preceq_{vl} \{\!\{1\}\!\} \preceq_{vl} \{\!\{2,3\}\!\} \preceq_{vl} \{\!\{2,3,3\}\!\}$ $\preceq_{vl} \{\!\{2,2,3\}\!\} \preceq_{vl} \{\!\{2,2,3,3\}\!\} \preceq_{vl} \{\!\{1,3\}\!\} \preceq_{vl} \{\!\{1,3,3\}\!\} \preceq_{vl} \{\!\{1,2\}\!\} \preceq_{vl} \{\!\{1,2,2\}\!\}$ $\preceq_{vl} \{\!\{1,2,3\}\!\} \preceq_{vl} \{\!\{1,2,3,3\}\!\} \preceq_{vl} \{\!\{1,2,2,3\}\!\} \preceq_{vl} \{\!\{1,2,2,3,3\}\!\}$
Length-variety-lex (LVL)	$\emptyset \preceq_{lvl} \{\!\{3\}\!\} \preceq_{lvl} \{\!\{2\}\!\} \preceq_{lvl} \{\!\{1\}\!\} \preceq_{lvl} \{\!\{3,3\}\!\} \preceq_{lvl} \{\!\{2,2\}\!\} \preceq_{lvl} \{\!\{2,3\}\!\} \preceq_{lvl} \{\!\{1,3\}\!\}$ $\preceq_{lvl} \{\!\{1,2\}\!\} \preceq_{lvl} \{\!\{2,3,3\}\!\} \preceq_{lvl} \{\!\{2,2,3\}\!\} \preceq_{lvl} \{\!\{1,3,3\}\!\} \preceq_{lvl} \{\!\{1,2,2\}\!\} \preceq_{lvl} \{\!\{1,2,3\}\!\}$ $\preceq_{lvl} \{\!\{2,2,3,3\}\!\} \preceq_{lvl} \{\!\{1,2,3,3\}\!\} \preceq_{lvl} \{\!\{1,2,2,3\}\!\} \preceq_{lvl} \{\!\{1,2,2,3,3\}\!\}$
Variety-length-lex (VLL)	$\emptyset \preceq_{vll} \{\!\{3\}\!\} \preceq_{vll} \{\!\{2\}\!\} \preceq_{vll} \{\!\{1\}\!\} \preceq_{vll} \{\!\{3,3\}\!\} \preceq_{vll} \{\!\{2,2\}\!\} \preceq_{vll} \{\!\{2,3\}\!\} \preceq_{vll} \{\!\{1,3\}\!\}$ $\preceq_{vll} \{\!\{1,2\}\!\} \preceq_{vll} \{\!\{2,3,3\}\!\} \preceq_{vll} \{\!\{2,2,3\}\!\} \preceq_{vll} \{\!\{1,3,3\}\!\} \preceq_{vll} \{\!\{1,2,2\}\!\}$ $\preceq_{vll} \{\!\{2,2,3,3\}\!\} \preceq_{vll} \{\!\{1,2,3\}\!\} \preceq_{vll} \{\!\{1,2,3,3\}\!\} \preceq_{vll} \{\!\{1,2,2,3\}\!\} \preceq_{vll} \{\!\{1,2,2,3,3\}\!\}$

by the vll-interval $\langle \{\!\{2,3\}\!\}, \{\!\{2,2,3,3\}\!\} \rangle_{vll}$, while the LVL, LL, or VL representations cannot.

Compactness

The notion of expressiveness concerns the exactness of the representation. However, a domain D of a multiset variable might not be exactly represented using any of the eight representations, i.e., $D \subset cl_\alpha(D)$. In such cases, $cl_\alpha(D)$ is an approximation that contains some undesired values, and our expressiveness notion does not apply. In this section, we define a new notion called *compactness* to compare the eight representations. This definition is based on a comparison of the size of the domains, and is different from the notion of dominance which is based on the size of search tree (Jefferson 2007).

Definition 8. *Given a universe U and two different multiset representations A and B. A is* as compact as *B if* $\forall S \subseteq U, |cl_A(S)| = |cl_B(S)|$. *A is* more compact *than B if* $\forall S \subseteq U, |cl_A(S)| \leq |cl_B(S)|$ *and* $\exists S \subseteq U, |cl_A(S)| < |cl_B(S)|$. *A and B are* compactly incomparable *if neither one of them is more compact than the other.*

The following proposition characterizes the compactness of the eight orderings.

Proposition 4. *(i) The LVL/LVC representation is more compact than the LL/LC representation and compactly incomparable to the VLL/VLC representation. (ii) The VLL/VLC representation is more compact than the VL/VC representation. (iii) The LL/LC representation is compactly incomparable to the VL/VC representation.*

In Table 1, suppose we want to represent the set S of all multisets whose variety is 2. Both the LVL and LL representations cannot exactly represent S and give a α-closure with the same lower and upper bounds (i.e., $\{\!\{2,3\}\!\}$ and $\{\!\{2,2,3,3\}\!\}$ respectively). Both lvl- and ll-intervals contain undesired values. By comparing their compactness, $|cl_{lvl}(S)| = 9 < |cl_{ll}(S)| = 10$. The LVL representation is more compact than LL representation.

Using the VL/VLL representations for multiset variables would be useful when we have tight constraints on the varieties of the multiset variables. For instance, Law, Lee, and Woo (2009) demonstrated the value of this on extended Steiner system problems in which there are tight constraints over the varieties. On the other hand, the LL/LVL representations would favour the kind of problems with more cardinality restrictions or with variables having fixed cardinalities.

Empirical Comparisons

Before we apply the eight representations to model and solve multiset problems, we first empirically evaluate their expressiveness and compactness. We perform experiments to compare the size of the eight representations of a set D of multisets when different cardinality and variety constraints are imposed. In the experiment, the universe U is a multiset which contains 10 occurrences of elements 1 to 5. For all instances, D is a randomly generated subset of the power set of U. The comparison aims at measuring the compactness of different representations in approximating D. We record $|cl_\alpha(D)|$, the number of multisets in the α-closure of D that satisfies the cardinality and the variety constraints, where α refers to the eight representations: LL, LC, VL, VC, LVL, LVC, VLL, and VLC. Due to space limitation, we summarize the observations as follows.

When both cardinality and variety are fixed, the LVL/LVC and VLL/VLC representations can always exactly represent the domain values, giving the corresponding minimal α-interval $cl_\alpha(S)$. For all instances, the LVL/LVC and VLL/VLC representations demonstrate a large reduction in the domain size when compared with the LL/LC and VL/VC representations.

When the variety is fixed, the VLL/VLC ordering first considers the variety of each multiset and narrows down the bounds to a larger extent by removing the multisets with unwanted varieties. For each variety, the multisets are then ordered by their cardinality, which allows further pruning of the multisets with undesired cardinalities on the domain bounds. Thus, the VLL/VLC representation can always give the exact representation and achieve on average one to two orders of magnitude reduction in the domain size when compared with the LL/LC and VL/VC representations. In contrast, the LVL/LVC representation can always give the exact representation when the cardinality is fixed.

When the cardinality and variety are constrained to certain ranges, although all eight representations fail to give the exact representation for all instances, the LVL/LVC and VLL/VLC representations are more compact than the

LL/LC and VL/VC representations respectively.

To conclude, the LVL/LVC and VLL/VLC representations are always more compact than the LL/LC and VL/VC respectively. This means that they will usually give tighter bounds during constraint propagation. In the following, we study how the eight representations behave in practice as bounds propagation in a multiset solver.

Bounds Consistency

Since a multiset domain is totally ordered in the eight representations, we can enforce bounds consistency. To be more precise, we define bounds consistency on a k-ary constraint on multiset variables (for any k).

Definition 9. *Bounds Consistency (BC)*
*Let $S_1, \ldots, S_n$ be multiset variables with interval domains $D(S_i) = \langle m_{S_i}, M_{S_i} \rangle$. Given a constraint $\mathcal{C}$ over $S_1, \ldots, S_n$ and an α ordering, a value m_i for variable S_i has an α-*bound support *$(m_1, \ldots, m_n)$ if the support satisfies $\mathcal{C}$ and $\forall m_i, m_{S_i} \preceq_\alpha m_i \preceq_\alpha M_{S_i}$.*

The constraint $\mathcal{C}$ is bounds consistent *iff for each S_i, both m_{S_i} and M_{S_i} have α-bound supports.*

The eight representations offer greater expressiveness, but we have to be careful that reasoning remains tractable. Indeed, even with a single unary constraint, we can get intractability.

Theorem 1. *There exists a constraint on one set variable such that enforcing BC on subset bounds is polynomial but enforcing BC on LL bounds is NP-hard.*

Proof. Reduction from 3-SAT with N variables, X_1 to X_N and M clauses. We construct a set variable S with elements that have the following meaning: $2i$ represents a truth assignment in which X_i is true whilst $2i-1$ represents a truth assignment in which X_i is false ($1 \le i \le N$), and each integer above $2N$ represents one of the (polynomial number of) distinct clauses. We consider an unary constraint on this set variable which is satisfied only when the set contains integers representing a proper truth assignment (that is, $2i \in$S iff $2i-1 \notin S$ for $1 \le i \le N$) and this assignment satisfies the clauses represented by the integers in the set greater than $2N$, or the set contains integers representing a superset of a proper truth assignment (that is, either $2i$ or $2i-1$ or both occur in S for $1 \le i \le N$). Subset bounds are polynomial to compute since, if the upper bound includes a proper truth assignment, we leave the upper bound untouched and adjust the lower bound to include any necessary elements in linear time and, where needed, check the truth assignment. On the other hand, if the upper bound does not include a proper truth assignment, the unary constraint has no support. By comparison, length-lex bounds are NP-hard to compute. We consider domains that fix the possible and necessary elements to be the clause that we wish to decide, and make none of the other integers necessary but all of them possible. Then, enforcing bound consistency on the length-lex bounds will allow us to decide the satisfiability of the original formula. □

It is worth noting that the opposite does not hold. If LL bounds are polynomial to compute, then subset bounds are too.

Theorem 2. *Given an n-ary constraint on set and/or multiset variables. If enforcing BC on LL bounds is polynomial, then enforcing BC on subset bounds is also polynomial.*

Proof. (sketch) Let the possible values of a set variable S be $\{1, \ldots, n\}$. We can convert subset bounds into LL bounds easily by ordering the sets first by cardinality and then lexicographically. This operation is polynomial. After enforcing BC on LL bounds, we can then convert LL bounds back to subset bounds using the inclusion propagator (Gervet and Van Hentenryck 2006). Such conversion is also polynomial. Thus, if enforcing BC on LL bounds is polynomial, then enforcing BC on subset bounds is also polynomial. □

With two unary constraints, Sellmann's Lemma 1 shows that finding the fixpoint on the LL representation of a single set variable is NP-hard (Sellmann 2009). Given the above theorems, enforcing BC on LL bounds is NP-hard. However, exponential-time propagation algorithms may still help reduce runtimes (Yip and Van Hentenryck 2010).

Here, we show an example on how BC works on the domains in the LL and LVL representations.

Given the universe $U = \{\!\{1,1,1,2,2,2,3,3,3\}\!\}$ and multiset variables X, Y, and Z. The constraints are: $|X| = |Y| = |Z| = 3$, $\|Z\| = 1$, and $X \cap Y = Z$. The initial domains are $D(X) = D(Y) = D(Z) = \langle \emptyset, \{\!\{1,1,1,2,2,2,3,3,3\}\!\} \rangle_{lvl}$. In LVL representation, enforcing $|X| = |Y| = |Z| = 3$ tightens the bounds to have cardinality 3, i.e., $D(X) = D(Y) = D(Z) = \langle \{\!\{3,3,3\}\!\}, \{\!\{1,2,3\}\!\} \rangle_{lvl}$. The bounds corresponds to the occurrence vectors $\langle 0,0,3 \rangle$ and $\langle 3,0,0 \rangle$. Since $\|\{\!\{1,2,3\}\!\}\| \neq 1$, the upper bound of Z is updated to $\{\!\{1,1,1\}\!\}$, resulting $D(Z) = \langle \{\!\{3,3,3\}\!\}, \{\!\{1,1,1\}\!\} \rangle_{lvl}$. This triggers the propagation on $X \cap Y = Z$ and tightens the upper bounds of X and Y. After constraint propagation, $X = Y = \langle \{\!\{3,3,3\}\!\}, \{\!\{1,1,1\}\!\} \rangle_{lvl}$. Now, the problem is bounds consistent and $|D(X)| = |D(Y)| = |D(Z)| = 3$. However, in the LL representation, the problem is bounds consistent after enforcing the cardinality constraint $|X| = |Y| = |Z| = 3$. $D(X) = D(Y) = D(Z) = \langle \{\!\{3,3,3\}\!\}, \{\!\{1,1,1\}\!\} \rangle_{ll}$ and $|D(X)| = |D(Y)| = |D(Z)| = 10$. Thus, different representations result in different domain size after enforcing BC, and LVL gives a tighter bound than LL in this example.

Experimental Results

To verify the feasibility and efficiency of our proposal, we adapt and simplify the implementation of the length-lex representation for set variables (Van Hentenryck et al. 2008) to implement the eight representations (LL, LVL, VL, VLL, LC, LVC, VC, VLC) for multiset variables in ILOG Solver 6.0 (ILOG 2003). We have also developed the ternary intersection ($X \cap Y = Z$) and unionplus ($X \uplus Y = Z$) multiset constraints, which are not available in the original LL implementation.

We perform experiments on the extended Steiner system and the generalized social golfer problem. They are

Table 2: Experimental results of the extended Steiner system.

	SB+CR+VR		LL		LVL		VL		VLL	
t,k,u,b,v	Fail	Time	Fail	Time	Fail	Time	Fail	Time	Fail	Time
2,4,5,4,2	57329	3.59	19187	1.48	**2930**	**0.34**	3790	95.37	2945	3.38
2,4,5,5,2	356785	28.71	89768	10.04	**19718**	**3.13**	30755	541.13	19991	14.32
3,4,4,4,2	1710	0.1	942	0.08	**278**	**0.03**	309	1.77	305	0.58
3,4,4,5,2	30034	2.36	13541	1.39	**658**	**0.11**	922	20.33	729	15.13
3,4,5,5,3	312397	22.17	38109	5.84	**12195**	**1.36**	-	-	12363	7.23
3,4,5,6,3	2108410	190.15	281911	57.83	**103163**	**13.39**	-	-	106145	63.83
3,4,5,7,3	9813128	1097	1352165	380.42	**384145**	**63.05**	-	-	398511	285.16

Table 3: Experimental results of the generalized social golfer problem.

	SB+CR+VR		LL		LVL		VL		VLL	
w,m,n,g,p,v	Fail	Time	Fail	Time	Fail	Time	Fail	Time	Fail	Time
3,3,3,2,4,2	14934	1.61	15108	0.94	14479	0.87	**2171**	0.44	2395	**0.27**
3,3,4,2,4,2	394570	40.29	111102	6.41	103756	5.59	**39**	0.06	**39**	**0.05**
3,3,4,2,5,2	185839	20.32	181801	12.37	172818	11.27	**11536**	8.61	12428	**2.84**
4,3,4,2,4,2	-	-	14071439	1003.03	12983736	874.96	**151132**	78.47	**151132**	**41.6**
4,3,4,2,5,2	-	-	12818684	1103	12496315	1046.14	**1035895**	437.89	1098395	**173.74**
3,4,3,2,4,3	2631024	348.04	1889782	129.28	1510939	94.21	**21**	**0.28**	**21**	0.29
3,4,4,2,4,3	-	-	4062535	280.02	3339400	210.61	**27**	3.99	**27**	**3.95**

run on a Sun Blade 2500 ($2 \times$ 1.6GHz US-IIIi) workstation with 2GB memory. We report the number of fails (i.e., the number of backtracks occurred in solving a model) and CPU time in seconds to find and prove the optimal solution for each instance. Comparisons are made among the subset bounds representations with cardinality-variety reasoning (SB+CR+VR) (Law, Lee, and Woo 2009) and the eight representations we have implemented. Since the results of the four colex representations (LC, LVC, VC, VLC) are similar to their corresponding lex counterparts (LL, LVL, VL, VLL), they are not reported in the tables. In the tables, the first column shows the problem instances. The subsequent columns show the results of using various representations. The best number of fails and CPU time among the results for each instance are highlighted in bold. A cell labeled with "-" denotes a timeout after 20 minutes.

The extended Steiner system $ES(t, k, u, b)$, an important and practical multiset problem in information retrieval (Johnson and Mendelsohn 1972; Bennett and Mendelsohn 1980; Park and Blake 2008), is a collection of b blocks. Each block is a k-element multiset drawn from a u-element set whose elements can be drawn multiple times. For every two blocks in the collection, the cardinality of their intersection must be smaller than t. We adapt the problem to become an optimization problem which maximizes the sum of the varieties of the multisets. To further increase difficulty, we constrain each multiset variable to have variety at least v.

The generalized social golfer problem $SG(w, m, n, g, p)$ extends the social golfer problem (prob010 in CSPLib (Gent and Walsh 1999)) from sets to multiset, in which we schedule m teams of n members to g groups of p golfers over w weeks. Each group contains golfers from different teams and they play against each other. To maximize the socialization, the number of times two teams meet with each other again is minimized. Similar to the extended Steiner system, each multiset variable is constrained to have variety at least v.

Tables 2 and 3 show the experimental results of the extended Steiner system and the generalized social golfer problem respectively. All the four lex representations give fewer number of fails and faster runtime than the SB+CR+VR (Law, Lee, and Woo 2009). This confirms that the lex representations take advantage of the cardinality and variety information to give tighter bounds than the SB+CR+VR.

In the extended Steiner system, the LVL representation always achieves the fewest number of fails. There is about a 95% reduction in the number of fails when compared to the SB+CR+VR. The LVL representation achieves fewer number of fails than the VLL representation because the problem has tighter constraints on the cardinalities than the varieties of the multiset variables.

When comparing the results between LL and LVL, the latter performs better. This is because in the LVL representation, the multisets are ordered according to their varieties under the same cardinality. When enforcing BC, the multisets with the same varieties can be pruned together when they violate the variety constraints. However, in the LL representation, these multisets are scattered over the ordering and we cannot remove all of them from the domain at the same time, thus resulting in a larger search tree and number of fails. Similarly, VLL performs better than VL.

The instances listed in Table 2 are all satisfiable. In our experiments, there are some unsatisfiable instances, in which the number of fails and runtime of LVL and VLL can be slightly larger than LL and VL respectively. We also tried to fix both cardinalities and varieties of the multiset variables. Since the multisets are ordered lexicographically under a fixed cardinality and variety, LVL and VLL give the same number of fails.

For the generalized social golfer problem, VL and VLL perform better than LL and LVL because the problem has tighter constraints on the varieties than the cardinalities of the multiset variables. Since there are much more constraints in the problem when compared to those in the extended Steiner system, the generalized social golfer problem is more complicated. We observe that the VL representation always achieves the fewest number of fails. However, the VLL representation has the fastest runtime because the extra prunings in the VL representation cannot compensate the overhead in finding new bounds of multiset variables.

Conclusion

We have proposed eight representations for multiset variables, which integrate together information about the cardinality, variety, and position in the (co)lexicographic ordering. We have made a detailed comparison of the expressiveness and compactness between the eight different representations. The LVL/LVC and VLL/VLC representations are always more expressive and more compact than the LL/LC and VL/VC representations. Compactness is a new notion which lets us compare inexact representations. We have also performed experiments on some benchmark problems. Experimental results confirm that LVL and VLL usually give tighter bounds during constraint propagation, resulting in smaller search trees and better runtimes. In some cases, LVL performs better, and sometimes VLL. It would be interesting to study if the two representations can be linked together so that we can take advantage of each representation.

Acknowledgments

We thank the anonymous referees for constructive comments. The work described in this paper was substantially supported by grants (CUHK413808 and CUHK413710) from the Research Grants Council of Hong Kong SAR. Toby Walsh is funded by the Australian Department of Broadband, Communications and the Digital Economy, the ARC, and the Asian Office of Aerospace Research and Development through grant AOARD-104123.

References

Bennett, F., and Mendelsohn, E. 1980. Extended (2, 4)-designs. *J. Comb. Theory, Ser. A* 29(1):74–86.

Gent, I. P., and Walsh, T. 1999. CSPLib: A benchmark library for constraints. In *Proc. of CP'99*, 480–481.

Gervet, C., and Van Hentenryck, P. 2006. Length-lex ordering for set CSPs. In *Proc. of AAAI'06*, 48–53.

Gervet, C. 1997. Interval propagation to reason about sets: Definition and implementation of a practical language. *Constraints* 1(3):191–244.

ILOG. 2003. *ILOG Solver 6.0 Reference Manual*.

Jefferson, C. 2007. *Representations in Constraint Programming*. Ph.D. Dissertation, University of York, York, UK.

Johnson, D., and Mendelsohn, N. 1972. Extended triple systems. *Aequat. Math.* 8(3):291–298.

Law, Y. C.; Lee, J. H. M.; and Woo, M. H. C. 2009. Variety reasoning for multiset constraint propagation. In *Proc. of IJCAI'09*, 552–558.

Park, E., and Blake, I. 2008. Construction of extended Steiner systems for information retrieval. *Rev. Mat. Complut.* 21(1):179–190.

Sadler, A., and Gervet, C. 2004. Hybrid set domains to strengthen constraint propagation and reduce symmetries. In *Proc. of CP'04*, 604–618.

Sadler, A., and Gervet, C. 2008. Enhancing set constraint solvers with lexicographic bounds. *J. Heuristics* 14(1):23–67.

Sellmann, M. 2009. On decomposing knapsack constraints for length-lex bounds consistency. In *Proc. of CP'09*, 762–770.

Van Hentenryck, P.; Yip, J.; Gervet, C.; and Dooms, G. 2008. Bound consistency for binary length-lex set constraints. In *Proc. of AAAI'08*, 375–380.

Walsh, T. 2003. Consistency and propagation with multiset constraints: A formal viewpoint. In *Proc. of CP'03*, 724–738.

Yip, J., and Van Hentenryck, P. 2010. Exponential propagation for set variables. In *Proc. of CP'10*, 499–513.