Geometric decision procedures and the VC dimension of linear arithmetic theories

Dmitry Chistikov[1], Christoph Haase[2] and **Alessio Mansutti**[2]

[1]University of Warwick and [2]University of Oxford, UK

Linear Integer Arithmetic (LIA, a.k.a. Presburger arithmetic)

The first-order theory of $\langle \mathbb{Z}, 0, 1, +, \leq \rangle$

"Every integer is either even or odd"

$$\forall \mathrm{x}\, \exists \mathrm{y} : \mathrm{x} = 2\mathrm{y} \vee \mathrm{x} = 2\mathrm{y} + 1$$

Why Linear Integer Arithmetic?

- Number theory is (highly) undecidable
- LIA is decidable [Presburger, '29]
- Wide range of applications in verification, program synthesis, compiler optimisation...
- Starting point of several algorithmic paradigms

Algorithmic paradigms for LIA

Quantifier elimination [Presburger, '29]

$$\exists x : \varphi_{\mathsf{QF}}(x, \mathbf{y}) \;\equiv\; \psi_{\mathsf{QF}}(\mathbf{y})$$

QF : quantifier-free

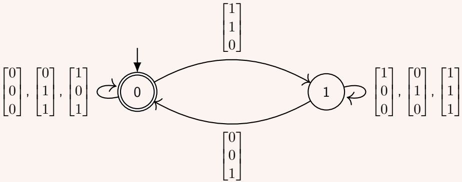
Algorithmic paradigms for LIA

Quantifier elimination [Presburger, '29]

$$\exists x : \varphi_{\mathsf{QF}}(x, \mathbf{y}) \;\equiv\; \psi_{\mathsf{QF}}(\mathbf{y}) \qquad\qquad \mathsf{QF} : \text{quantifier-free}$$

Automata-based procedures [Büchi, '60]

$$z = x + y \quad \rightsquigarrow \quad \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{x+y} \end{bmatrix} :$$



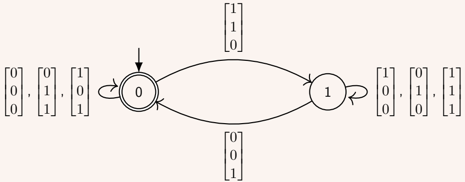
Algorithmic paradigms for LIA

Quantifier elimination [Presburger, '29]

$$\exists x : \varphi_{\mathsf{QF}}(x, \mathbf{y}) \;\equiv\; \psi_{\mathsf{QF}}(\mathbf{y}) \qquad\qquad \mathsf{QF} : \text{quantifier-free}$$

Automata-based procedures [Büchi, '60]

$$z = x + y \quad \rightsquigarrow \quad \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{x+y} \end{bmatrix}:$$



Geometric procedures (semilinear sets) [Ginsburg and Spanier, '66]

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{x+y} \end{bmatrix}: \qquad \begin{bmatrix}0\\0\\0\end{bmatrix} + \begin{bmatrix}1\\0\\1\end{bmatrix} \cdot \mathbb{N} + \begin{bmatrix}0\\1\\1\end{bmatrix} \cdot \mathbb{N}$$

Algorithmic paradigms for LIA

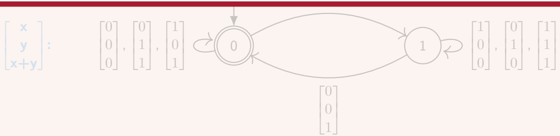
Quantifier elimination [Presburger, '29]

Problem:

- Quantifier elimination and automata-based procedures are optimal for deterministic time (3EXPTIME);
- geometric procedures are (currently) not!

Automata

$z = x + y \quad \rightsquigarrow$



Geometric procedures (semilinear sets) [Ginsburg and Spanier, '66]

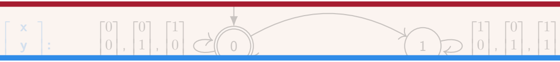
$$\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \\ \mathsf{x+y} \end{bmatrix} : \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \cdot \mathbb{N} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \cdot \mathbb{N}$$

Algorithmic paradigms for LIA

Quantifier elimination [Presburger, '29]

Automata

$z = x + y \rightsquigarrow$



Geometric

Problem:

- Quantifier elimination and automata-based procedures are optimal for deterministic time ($3\textsc{ExpTime}$);
- geometric procedures are (currently) not!

In this work:

- We give the first **optimal geometric procedure** for LIA...
- ...and from it characterise the **VC dimension** of the theory *(resolves a conjecture by Nguyen and Pak [Comb., '19])*.
- Analogous results for Linear Real Arithmetic.

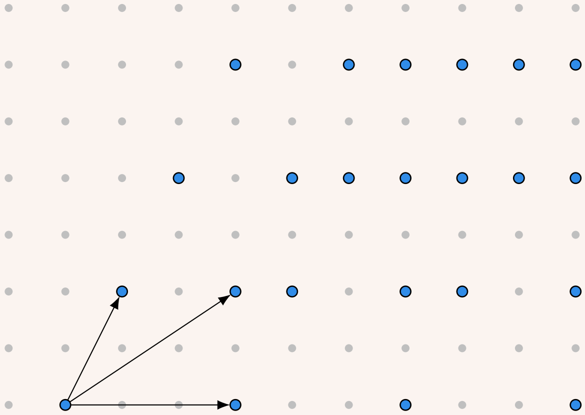
Semilinear sets

Arithmetic progression

$b + i \cdot p$, where $i \in \mathbb{N}$

b base point, p period

Semilinear sets



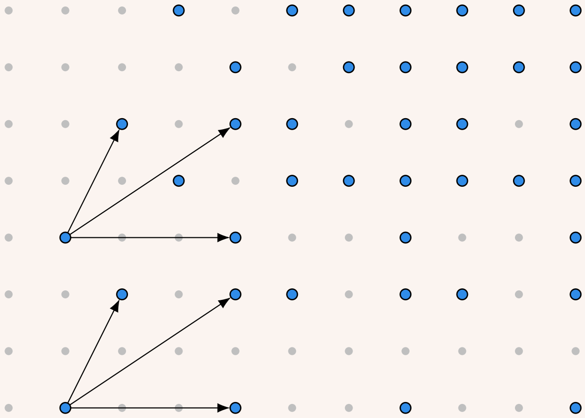
Linear set

(arithmetic progression
in multiple dimensions)

$L(\mathbf{b}, P)$

$\mathbf{b}$ base point, P periods

Semilinear sets



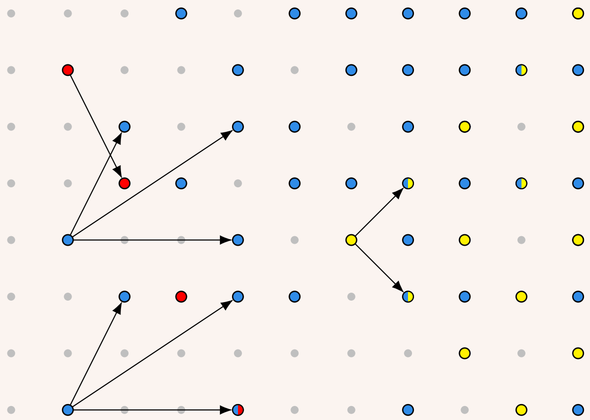
Hybrid linear set

(finite union of linear sets having the same periodic behaviour)

$L(B, P)$

B bases, P periods

Semilinear sets

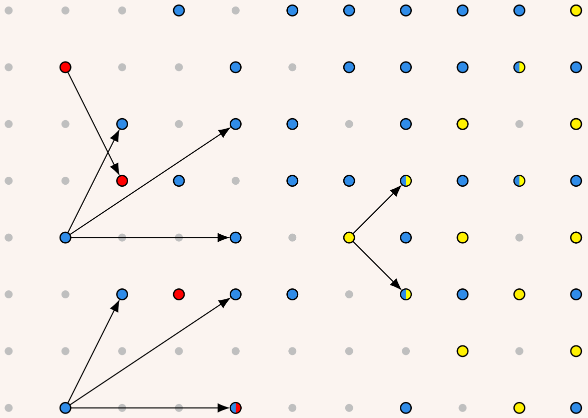


Semilinear set
(finite union of
hybrid linear sets)

$$\bigcup_{i\in I} L(B_i, P_i)$$

I index, B_i bases, P_i periods

Semilinear sets

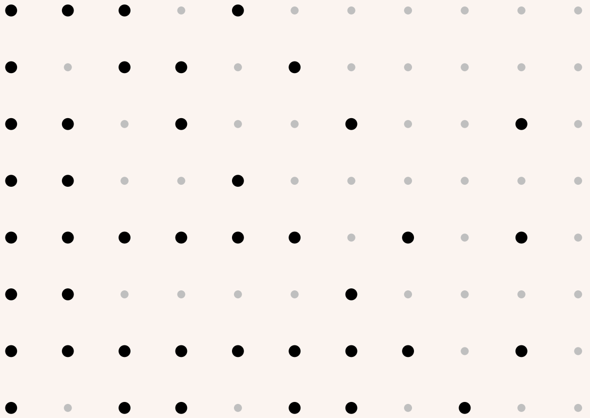


Ginsburg & Spanier, '66

The set of solutions of a system of linear inequalities over $\mathbb{Z}$ is semilinear. Semilinear sets are closed under

- union
- projection
- complementation

Semilinear sets



Ginsburg & Spanier, '66

The set of solutions of a system of linear inequalities over $\mathbb{Z}$ is semilinear. Semilinear sets are closed under

- union
- projection
- complementation

Main problem: How to compute the complement of a semilinear set optimally?

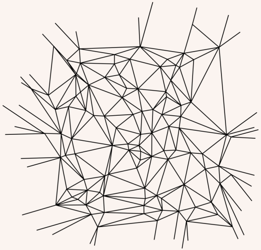
Complementation of a semilinear set

Input: $M = \bigcup_{i\in I} L(B_i, P_i) \subseteq \mathbb{Z}^d$

Output: $\{(C_j, Q_j)\}_{j\in J}$ such that $\mathbb{Z}^d \setminus M = \bigcup_{j \in J} L(C_j, Q_j)$.

Briefly,

1. we compute a **triangulation** of $\mathbb{Z}^d$ in generalised simplices of dimension $\leq d$...
2. ...chosen so that complementing M inside each simplex is simple.

(1) 

generalised simplex of dimension $\leq d$:

$(\operatorname{conv} V + \operatorname{cone} W) \cap \mathbb{Z}^d$ where $\#V + \#W \leq d+1$.

It is an hybrid linear set with periods $W \subseteq \mathbb{Z}^d$.

(2) our algorithm uses the least amount of period vectors to describe the simplices; they are the only periods needed to construct $\mathbb{Z}^d \setminus M$.

Complementation of a semilinear set

Input: $M = \bigcup_{i \in I} L(B_i, P_i) \subseteq \mathbb{Z}^d$

Output: $\{(C_j, Q_j)\}_{j \in J}$ such that $\mathbb{Z}^d \setminus M = \bigcup_{j \in J} L(C_j, Q_j)$.

Briefly

Theorem 1 (A geometric procedure for LIA)

Let $\Phi(\mathbf{x})$ in LIA. Then, $\{\mathbf{x} : \Phi(\mathbf{x})\} = \bigcup_{i \in I} L(B_i, P_i)$ where

- *$\|B_i\|$ and $\|P_i\|$ are triply exponential in $|\Phi|$ (doubly exponential bitsize)*
- *$\#I$ is doubly exponential in $|\Phi|$; vectors in each P_i are linearly independent.*

The family $\{(B_i, P_i)\}_{i \in I}$ can be computed in 3EXPTIME in $|\Phi|$.

(1) 

$(\operatorname{conv} V + \operatorname{cone} W) \cap \mathbb{Z}^d$ where $\#V + \#W \leq d+1$.

It is an hybrid linear set with periods $W \subseteq \mathbb{Z}^d$.

(2) our algorithm uses the least amount of period vectors to describe the simplices; they are the only periods needed to construct $\mathbb{Z}^d \setminus M$.

Complementation of a semilinear set

Input: $M = \bigcup_{i \in I} L(B_i, P_i) \subseteq \mathbb{Z}^d$

Output: $\{(C_j, Q_j)\}_{j \in J}$ such that $\mathbb{Z}^d \setminus M = \bigcup_{j \in J} L(C_j, Q_j)$.

Briefly

Theorem 1 (A geometric procedure for LIA)

Let $\Phi(\mathbf{x})$ in LIA. Then, $\{\mathbf{x} : \Phi(\mathbf{x})\} = \bigcup_{i \in I} L(B_i, P_i)$ where

- *$\|B_i\|$ and $\|P_i\|$ are triply exponential in $|\Phi|$ (doubly exponential bitsize)*
- *$\#I$ is doubly exponential in $|\Phi|$; vectors in each P_i are linearly independent.*

The family $\{(B_i, P_i)\}_{i \in I}$ can be computed in $3\mathrm{EXPTIME}$ in $|\Phi|$.

(1) 

$(\mathrm{conv}\, V + \mathrm{cone}\, W) \cap \mathbb{Z}^d$ where $\#V + \#W \leq d+1$.

It is an hybrid linear set with periods $W \subseteq \mathbb{Z}^d$.

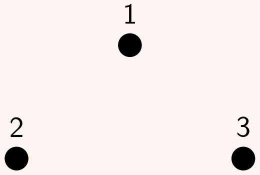
(2) our algorithm uses the least amount of period vectors to describe the simplices; they are the only periods needed to construct $\mathbb{Z}^d \setminus M$.

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.

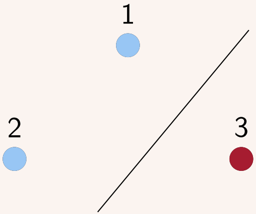


$S = \{1, 2, 3\}$

In how many ways we can classify these three points using one straight line?

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



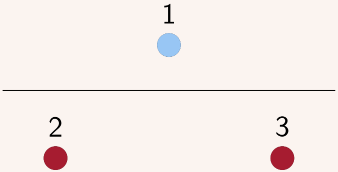
$S = \{1, 2, 3\}$

In how many ways we can classify these three points using one straight line?

subsets found: $\{1, 2\}$

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



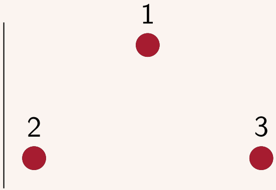
$S = \{1, 2, 3\}$

In how many ways we can classify these three points using one straight line?

subsets found: $\{1, 2\}$, $\{1\}$

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



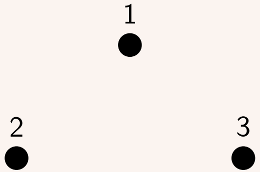
$S = \{1, 2, 3\}$

In how many ways we can classify these three points using one straight line?

subsets found: $\{1, 2\}$, $\{1\}$, $\emptyset$

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



$S = \{1, 2, 3\}$

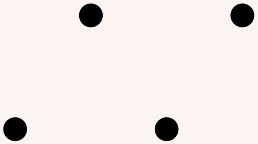
In how many ways we can classify these three points using one straight line?

2^3

subsets found: $\{1, 2\}$, $\{1\}$, $\emptyset$, ...

Vapnik–Chervonenkis (VC) dimension

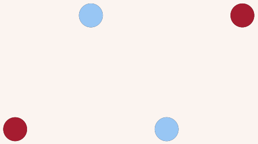
- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



In how many ways we can classify four abritrary points using one straight line?

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.

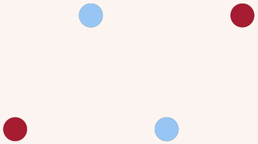


In how many ways we can classify four abritrary points using one straight line?

$$< 2^4$$

Vapnik–Chervonenkis (VC) dimension

- Measure of the capacity ($\sim$ expressiveness) of a set of functions that can be learned by a binary classification model. It is used to bound the sample complexity.
- Roughly speaking, it is the maximal cardinality of a set S for which all elements in 2^S can be classified by opportunely changing the parameters of a classifier.



In how many ways we can classify four abritrary points using one straight line?

$< 2^4$

The VC dimension of a straight-line classifier is 3.

The VC dimension of LIA

VC dimension applies to formulae $\Phi(\mathbf{x}, \mathbf{y})$ seen as classifiers having $\mathbf{y}$ as parameters.

The VC dimension of LIA

VC dimension applies to formulae $\Phi(\mathbf{x}, \mathbf{y})$ seen as classifiers having $\mathbf{y}$ as parameters.

Nguyen and Pak, *[Combinatorica '19]*:

conjecture that a doubly exponential upper bound on $\mathrm{VC}(F)$ holds in the general setting. It is unlikely that such an upper bound could be established by straightforward quantifier elimination, which generally results in triply exponential blow up (see [Wei97, Thm 3.1]).

The VC dimension of LIA

VC dimension applies to formulae $\Phi(\mathbf{x}, \mathbf{y})$ seen as classifiers having $\mathbf{y}$ as parameters.

Nguyen and Pak, *[Combinatorica '19]*:

> conjecture that a doubly exponential upper bound on $\mathrm{VC}(F)$ holds in the general setting. It is unlikely that such an upper bound could be established by straightforward quantifier elimination, which generally results in triply exponential blow up (see [Wei97, Thm 3.1]).

The doubly exponential bound on $\#I$ in our previous theorem does not show up in any form in the quantifier elimination procedure. Our geometric procedure shows that:

Theorem 2

Let $\Phi(\boldsymbol{x}, \boldsymbol{y})$ in LIA. Its VC dimension is doubly exponential in $|\Phi|$.

Conclusion

We define the first **optimal algorithm for complementing a semilinear set**, which

1. gives us a geometric procedure to decide LIA in $3\textsc{ExpTime}$ (as QE or automata)
2. shows that LIA has a doubly exponential VC dimension.

These results are obtained by extending similar results over the reals. We define

- a geometric procedure to decide Linear Real Arithmetic (LRA) in $2\textsc{ExpTime}$...
- ...from which we deduce that LRA has an exponential VC dimension.