Data compression and definability of types in stable and dependent formulas

Chris Laskowski

University of Maryland

Paris, 26 July, 2010

"Original" Compression schemes

Suppose $\mathscr{C} \subseteq {}^X 2$ is a set of 'concepts'.
Let $\mathscr{C}|_{\mathrm{fin}} = \{c|Y : c \in \mathscr{C} \text{ and } Y \subseteq X, Y \text{ finite}\}$
and $\mathscr{C}|_{\leq d} = \{c|Z : c \in \mathscr{C} \text{ and } Z \subseteq X, |Z| \leq d\}$.

Definition (Littlestone-Warmuth, 1986)

A *d*-dimensional compression scheme consists of a *compression function* $\kappa : \mathscr{C}|_{\mathrm{fin}} \to \mathscr{C}|_{\leq d}$ and a *reconstruction function* $\rho : \mathscr{C}|_{\leq d} \to {}^X 2$ satisfying

$$\kappa(c|Y) \subseteq c|Y \subseteq \rho(\kappa(c|Y))$$

for all $c \in \mathscr{C}$ and finite $Y \subseteq X$.

"Original" Compression schemes

Suppose $\mathscr{C} \subseteq {}^X 2$ is a set of 'concepts'.
Let $\mathscr{C}|_{\mathrm{fin}} = \{c|Y : c \in \mathscr{C} \text{ and } Y \subseteq X, Y \text{ finite}\}$
and $\mathscr{C}|_{\leq d} = \{c|Z : c \in \mathscr{C} \text{ and } Z \subseteq X, |Z| \leq d\}$.

Definition (Littlestone-Warmuth, 1986)

A *d-dimensional compression scheme* consists of a *compression function* $\kappa : \mathscr{C}|_{\mathrm{fin}} \to \mathscr{C}|_{\leq d}$ and a *reconstruction function* $\rho : \mathscr{C}|_{\leq d} \to {}^X 2$ satisfying

$$\kappa(c|Y) \subseteq c|Y \subseteq \rho(\kappa(c|Y))$$

for all $c \in \mathscr{C}$ and finite $Y \subseteq X$.

Open Question Does every d-dimensional VC class $\mathscr{C}$ of concepts have a d-dimensional compression scheme?

"Original" Compression schemes

Suppose $\mathscr{C} \subseteq {}^X 2$ is a set of 'concepts'.
Let $\mathscr{C}|_{\mathrm{fin}} = \{c|Y : c \in \mathscr{C} \text{ and } Y \subseteq X, Y \text{ finite}\}$
and $\mathscr{C}|_{\leq d} = \{c|Z : c \in \mathscr{C} \text{ and } Z \subseteq X, |Z| \leq d\}$.

Definition (Littlestone-Warmuth, 1986)

A *d-dimensional compression scheme* consists of a *compression function* $\kappa : \mathscr{C}|_{\mathrm{fin}} \to \mathscr{C}|_{\leq d}$ and a *reconstruction function* $\rho : \mathscr{C}|_{\leq d} \to {}^X 2$ satisfying

$$\kappa(c|Y) \subseteq c|Y \subseteq \rho(\kappa(c|Y))$$

for all $c \in \mathscr{C}$ and finite $Y \subseteq X$.

Open Question Does every d-dimensional VC class $\mathscr{C}$ of concepts have a d-dimensional compression scheme?
Warmuth has offered a \$600 prize for an answer in either direction.

Extended Compression schemes

To get a better behaved notion, allow finitely many reconstruction functions.

Definition

Fix $\mathscr{C} \subseteq {}^X 2$. A *$d$-dimensional extended compression scheme* consists of a compression function $\kappa : \mathscr{C}|_{\mathrm{fin}} \to X^d$ and finitely many reconstruction functions $\rho_i : X^d \to {}^X 2$ such that for every $c \in \mathscr{C}$ and $Y \subseteq_{fin} X$,

- $\mathrm{range}(\kappa(c|Y)) \subseteq Y$ and
- $\rho_i(\kappa(c|Y))$ extends $c|Y$ for at least one i.

This is equivalent to definitions proposed and studied by Litman-Ben-David, Basu, and Floyd-Warmuth.

Question: Which concept classes $\mathscr{C} \subseteq {}^X 2$ have d-dimensional extended compression schemes?

Question: Which concept classes $\mathscr{C} \subseteq {}^X 2$ have d-dimensional extended compression schemes?

- If X is finite, then all $\mathscr{C} \subseteq {}^X 2$ do.

Question: Which concept classes $\mathscr{C} \subseteq {}^X2$ have d-dimensional extended compression schemes?

• If X is finite, then all $\mathscr{C} \subseteq {}^X2$ do.

• If X is infinite and $\mathscr{C}$ has a d-dimensional extended compression scheme (with k reconstruction functions), then for $Y \subseteq X$ finite, elements of $\mathscr{C}_Y = \{c|Y : c \in \mathscr{C}\}$ are determined by $\kappa(c|Y) \in Y^d$ and by the choice of ρ_i. Thus, $|\mathscr{C}_Y| \leq k|Y|^d$.

Question: Which concept classes $\mathscr{C} \subseteq {}^X 2$ have d-dimensional extended compression schemes?

• If X is finite, then all $\mathscr{C} \subseteq {}^X 2$ do.

• If X is infinite and $\mathscr{C}$ has a d-dimensional extended compression scheme (with k reconstruction functions), then for $Y \subseteq X$ finite, elements of $\mathscr{C}_Y = \{c|Y : c \in \mathscr{C}\}$ are determined by $\kappa(c|Y) \in Y^d$ and by the choice of ρ_i. Thus, $|\mathscr{C}_Y| \leq k|Y|^d$.
It follows that $\mathscr{C}$ is a Vapnik-Chervonenkis (VC) class, i.e., for some m, no m-element subset of X is shattered by $\mathscr{C}$.

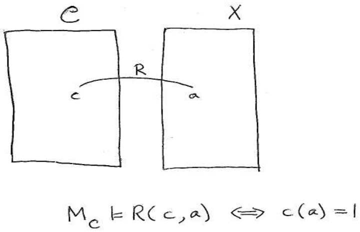
Which concept classes have extended compression schemes?

Which concept classes have extended compression schemes?

This is a model theoretic question!

Which concept classes have extended compression schemes?

This is a model theoretic question!
Given $\mathscr{C} \subseteq {}^X 2$, form a structure $M_{\mathscr{C}} = (\mathscr{C}, X, R(x, y))$.



Fact

If $\mathscr{C} \subseteq {}^X 2$ is given and the relation $R(x, y)$ is stable in the associated structure $M_{\mathscr{C}}$, then $\mathscr{C}$ has an extended compression scheme.

Fact

If $\mathscr{C} \subseteq {}^X 2$ is given and the relation $R(x, y)$ is stable in the associated structure $M_{\mathscr{C}}$, then $\mathscr{C}$ has an extended compression scheme.

Pf: Definability of types!

Fact

If $\mathscr{C} \subseteq {}^X 2$ is given and the relation $R(x, y)$ is stable in the associated structure $M_{\mathscr{C}}$, then $\mathscr{C}$ has an extended compression scheme.

Pf: Definability of types!

There is a formula $\psi(y, z_1, \ldots, z_d)$ such that for any $Y \subseteq X$ and for any $c \in \mathscr{C}$, there are $(b_1, \ldots, b_d) \in Y^d$ such that $R(c, Y) = \psi(Y, b_1, \ldots, b_d)$.

Fact

If $\mathscr{C} \subseteq {}^X 2$ is given and the relation $R(x, y)$ is stable in the associated structure $M_{\mathscr{C}}$, then $\mathscr{C}$ has an extended compression scheme.

Pf: Definability of types!

There is a formula $\psi(y, z_1, \ldots, z_d)$ such that for any $Y \subseteq X$ and for any $c \in \mathscr{C}$, there are $(b_1, \ldots, b_d) \in Y^d$ such that $R(c, Y) = \psi(Y, b_1, \ldots, b_d)$.

Compress via $\kappa(c|Y) = (b_1, \ldots, b_d)$ and reconstruct by $\rho(b_1, \ldots, b_d) = \psi(X, b_1, \ldots, b_d)$.

Question: If $\varphi(x, y)$ is stable, can we bound the d in a uniform defining formula $\psi(y, z_1, \ldots, z_d)$?

Question: If $\varphi(x, y)$ is stable, can we bound the d in a uniform defining formula $\psi(y, z_1, \ldots, z_d)$?

Answer: YES.

Question: If $\varphi(x, y)$ is stable, can we bound the d in a uniform defining formula $\psi(y, z_1, \ldots, z_d)$?

Answer: YES. $d \leq R_\varphi(x = x, 2)$.

Why? Recall $R_\varphi(\theta(x),2) \geq 0$ iff $\theta(x)$ is consistent and $R_\varphi(\theta(x),2) \geq n+1$ iff for some a, both $R_\varphi(\theta \wedge \varphi(x,a),2) \geq n$ and $R_\varphi(\theta \wedge \neg\varphi(x,a),2) \geq n$.

Why? Recall $R_\varphi(\theta(x),2) \geq 0$ iff $\theta(x)$ is consistent and $R_\varphi(\theta(x),2) \geq n+1$ iff for some a, both $R_\varphi(\theta \wedge \varphi(x,a),2) \geq n$ and $R_\varphi(\theta \wedge \neg\varphi(x,a),2) \geq n$.

Thus: • $\varphi(x,y)$ is stable iff $R_\varphi(x=x,2)$ is finite;
• $\{e : R_\varphi(\theta(x,e),2) \geq n\}$ is definable;
• If $R_\varphi(\theta,2) = n$, then for any a, **at most one** of $\theta \wedge \varphi(x,a)$, $\theta \wedge \neg\varphi(x,a)$ has $R_\varphi = n$.

Given $p \in S_\varphi(A)$, call a subtype $p_i \subseteq p$ one-element minimal if $R_\varphi(q, 2) = R_\varphi(p_i, 2)$ for all $p_i \subseteq q \subseteq p$ with $|q \setminus p_i| = 1$.

Given $p \in S_\varphi(A)$, call a subtype $p_i \subseteq p$ one-element minimal if $R_\varphi(q, 2) = R_\varphi(p_i, 2)$ for all $p_i \subseteq q \subseteq p$ with $|q \setminus p_i| = 1$.

Note: For any $p \in S_\varphi(A)$ there is a one-element minimal $p_i \subseteq p$ with $|p_i| \leq R_\varphi(x = x, 2)$.
Why? Let $p_0 = \emptyset$ and given p_i, let $p_{i+1} \subseteq p$ be any one-element extension of p_i of smaller 2-rank (if one exists).

Given $p \in S_\varphi(A)$, call a subtype $p_i \subseteq p$ one-element minimal if $R_\varphi(q,2) = R_\varphi(p_i,2)$ for all $p_i \subseteq q \subseteq p$ with $|q \setminus p_i| = 1$.

Note: For any $p \in S_\varphi(A)$ there is a one-element minimal $p_i \subseteq p$ with $|p_i| \leq R_\varphi(x = x, 2)$.
Why? Let $p_0 = \emptyset$ and given p_i, let $p_{i+1} \subseteq p$ be any one-element extension of p_i of smaller 2-rank (if one exists).

Check: For any $p \in S_\varphi(A)$, if $p_i \subseteq p$ is one-element minimal then p is defined by the formula "$R_\varphi(p_i \wedge \varphi(x,y),2) = R_\varphi(p_i,2)$."
Why? For $a \in A$, $\varphi(x,a) \in p \Rightarrow R_\varphi(p_i \wedge \varphi(x,a),2) = R_\varphi(p_i,2)$ by minimality of p_i and
$\varphi(x,a) \notin p \Rightarrow \neg\varphi(x,a) \in p \Rightarrow R_\varphi(p_i \wedge \neg\varphi(x,a),2) = R_\varphi(p_i,2) \Rightarrow$
$R_\varphi(p_i \wedge \varphi(x,a),2) \neq R_\varphi(p_i,2)$.

Caution: Even though every φ-type has a definition $\psi(y, z_1, \ldots, z_d)$ with $d \leq R_\varphi(x = x, 2)$, this does not imply that one can bound the size of a subtype $p_0 \subseteq p$ such that $R_\varphi(p_0, 2) = R_\varphi(p, 2)$.

A new notion:

A new notion:

Definition

A formula $\varphi(x, y)$ has **Uniform Definability Types over Finite Sets** (UDTFS) if there is a formula $\psi(y, z_1, \ldots, z_d)$ such that for every finite set A, $|A| \geq 2$ and every $p \in S_\varphi(A)$, there are $(b_1, \ldots, b_d) \in A^d$ such that

$$\varphi(x, a) \in p \iff \models \psi(a, b_1, \ldots, b_d)$$

for every $a \in A$.

Observation

If $\varphi(x, y)$ has UDTFS, then the uniformly definable family $\mathscr{C}_{\varphi(x,y)} = \{\varphi(c, M) : c \in Sort(x)\}$ has an extended compression scheme.

Which formulas have UDTFS?

- If $\varphi(x, y)$ is stable, then $\varphi(x, y)$ has UDTFS.

Which formulas have UDTFS?

- If $\varphi(x, y)$ is stable, then $\varphi(x, y)$ has UDTFS.

- If $\varphi(x, y)$ has UDTFS via $\psi(y, z_1, \dots, z_d)$, then for any finite set Y, $|S_\varphi(Y)| \leq |Y|^d$, so $\varphi(x, y)$ is dependent (NIP) with independence dimension at most d.

Which formulas have UDTFS?

- If $\varphi(x, y)$ is stable, then $\varphi(x, y)$ has UDTFS.

- If $\varphi(x, y)$ has UDTFS via $\psi(y, z_1, \ldots, z_d)$, then for any finite set Y, $|S_\varphi(Y)| \leq |Y|^d$, so $\varphi(x, y)$ is dependent (NIP) with independence dimension at most d.

Open Question Does every dependent formula have UDTFS?

Which formulas have UDTFS?

- If $\varphi(x,y)$ is stable, then $\varphi(x,y)$ has UDTFS.

- If $\varphi(x,y)$ has UDTFS via $\psi(y,z_1,\ldots,z_d)$, then for any finite set Y, $|S_\varphi(Y)| \leq |Y|^d$, so $\varphi(x,y)$ is dependent (NIP) with independence dimension at most d.

Open Question Does every dependent formula have UDTFS?
If you can prove this, you can petition Warmuth for $600.

Definability over Indiscernible Sequences

A plausibility argument:

A plausibility argument:

Theorem

A partitioned formula $\varphi(x, y)$ is stable if and only if there exists a formula $\psi(y, \overline{z})$ so that for all order indiscernible sequences A and all $p \in S_\varphi(A)$, there exists $\overline{a} \in A^d$ so that $\psi(y, \overline{a})$ defines p.

Theorem

A partitioned formula $\varphi(x, y)$ is dependent iff there exists a formula $\psi(y, \overline{z})$ so that for all finite order indiscernible sequences A and all $p \in S_\varphi(A)$ there exists $\overline{a} \in A^d$ so that $\psi(y, \overline{a})$ defines p.

The class of UDTFS formulas is well behaved:

The class of UDTFS formulas is well behaved:

• Closed under boolean combinations: If $\varphi(x, y)$ and $\psi(x, z)$ are both UDTFS, then so are $\neg\varphi(x, y)$ and $[\varphi \wedge \psi](x, yz)$.

The class of UDTFS formulas is well behaved:

• Closed under boolean combinations: If $\varphi(x,y)$ and $\psi(x,z)$ are both UDTFS, then so are $\neg\varphi(x,y)$ and $[\varphi \wedge \psi](x,yz)$.

• "Finitely many defining formulas suffice" Given $\varphi(x,y)$, if there are finitely many $\psi_i(y,z_1,\dots,z_d)$ such that for every finite A, every $p \in S_\varphi(A)$ is definable by some $\psi_i(y,a_1,\dots,a_d)$, then φ has UDTFS.

The class of UDTFS formulas is well behaved:

• Closed under boolean combinations: If $\varphi(x, y)$ and $\psi(x, z)$ are both UDTFS, then so are $\neg\varphi(x, y)$ and $[\varphi \wedge \psi](x, yz)$.

• "Finitely many defining formulas suffice" Given $\varphi(x, y)$, if there are finitely many $\psi_i(y, z_1, \ldots, z_d)$ such that for every finite A, every $p \in S_\varphi(A)$ is definable by some $\psi_i(y, a_1, \ldots, a_d)$, then φ has UDTFS.

• "Sufficiency of a single variable" [Guingona] If every formula $\varphi(x, \overline{y})$ with a single x-variable has UDTFS, then every formula $\varphi(\overline{x}, \overline{z})$ has UDTFS.

Theorem (H. Johnson-L, 2008)

If T is o-minimal then every formula $\varphi(\overline{x}, \overline{y})$ is UDTFS. It follows that the uniformly definable family $\mathscr{C}_{\varphi(\overline{x}, \overline{y})}$ has a d-dimensional extended compression scheme where $d = lg(\overline{x})$.

Theorem (H. Johnson-L, 2008)

If T is o-minimal then every formula $\varphi(\overline{x}, \overline{y})$ is UDTFS. It follows that the uniformly definable family $\mathscr{C}_{\varphi(\overline{x},\overline{y})}$ has a d-dimensional extended compression scheme where $d = lg(\overline{x})$.

In some sense, this was proved by Marker-Steinhorn who established definability of types for o-minimal structures with Dedekind complete order types.

Vincent Guingona's results:

Vincent Guingona's results:

- If T is weakly o-minimal, then every formula has UDTFS.

Vincent Guingona's results:

- If T is weakly o-minimal, then every formula has UDTFS.
- If φ has independence dimension one, then φ has UDTFS.

Vincent Guingona's results:

- If T is weakly o-minimal, then every formula has UDTFS.
- If φ has independence dimension one, then φ has UDTFS.
- If T is VC-minimal, then every formula has UDTFS.

Vincent Guingona's results:

- If T is weakly o-minimal, then every formula has UDTFS.
- If φ has independence dimension one, then φ has UDTFS.
- If T is VC-minimal, then every formula has UDTFS.
- If φ has density one, i.e., there is a constant k so that $|S_\varphi(A)| \leq k|A|$ for all finite sets A in the sort of y, then φ has UDTFS.

Some deeper results (also proved by Guingona):

Some deeper results (also proved by Guingona):

Theorem (Guingona)

Suppose there is an n such that for any set A of size n (in the sort of y), $|S_\varphi(A)| \leq \binom{n}{2} + \binom{n}{1}$ then φ has UDTFS.

Some deeper results (also proved by Guingona):

Theorem (Guingona)

Suppose there is an n such that for any set A of size n (in the sort of y), $|S_\varphi(A)| \leq \binom{n}{2} + \binom{n}{1}$ then φ has UDTFS.

Remark: If the independence dimension of φ is 2, then $|S_\varphi(A)| \leq \binom{n}{2} + \binom{n}{1} + 1$ by Sauer's theorem.

An *ict*-pattern with two rows consists of two formulas $\varphi(x, y)$ and $\psi(x, z)$ such that for every N there exist $\{b_i : i < N\}$ and $\{c_j : j < N\}$ such that each of the N^2 formulas

$$\varphi(x, b_{i^*}) \wedge \psi(x, c_{j^*}) \wedge \bigwedge_{i \neq i^*} \neg\varphi(x, b_i) \wedge \bigwedge_{j \neq j^*} \neg\psi(x, c_j)$$

indexed by $(i^*, j^*) \in N^2$ is consistent.

An *ict*-pattern with two rows consists of two formulas $\varphi(x,y)$ and $\psi(x,z)$ such that for every N there exist $\{b_i : i < N\}$ and $\{c_j : j < N\}$ such that each of the N^2 formulas

$$\varphi(x,b_{i^*}) \wedge \psi(x,c_{j^*}) \wedge \bigwedge_{i \neq i^*} \neg\varphi(x,b_i) \wedge \bigwedge_{j \neq j^*} \neg\psi(x,c_j)$$

indexed by $(i^*,j^*) \in N^2$ is consistent.

A theory T is *dp*-minimal if it does not admit an *ict*-pattern with two rows.

An *ict-pattern with two rows* consists of two formulas $\varphi(x, y)$ and $\psi(x, z)$ such that for every N there exist $\{b_i : i < N\}$ and $\{c_j : j < N\}$ such that each of the N^2 formulas

$$\varphi(x, b_{i^*}) \wedge \psi(x, c_{j^*}) \wedge \bigwedge_{i \neq i^*} \neg\varphi(x, b_i) \wedge \bigwedge_{j \neq j^*} \neg\psi(x, c_j)$$

indexed by $(i^*, j^*) \in N^2$ is consistent.

A theory T is *dp-minimal* if it does not admit an *ict*-pattern with two rows.

Theorem (Guingona)

If T is dp-minimal then every formula has UDTFS.

Bibliography

- S. Ben-David and A. Litman, Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes, *Discrete applied math*, vol 86(1) 3-25 (1998).
- S. Floyd and M. Warmuth, Sample compression, learnability and Vapnik-Chervonekis dimension, *Machine Learning*, vol 21(3), 269-304 (1995).
- V. Guingona, On uniform definability of types over finite sets, arXiv:1005.4924 and submitted to the *JSL*.
- H.R. Johnson and M.C. Laskowski, Compression schemes, stable definable families, and o-minimal structures, *Discrete and Computational Geometry* vol 43, 914-926 (2010).