Data Augmentation for Mathematical Objects

Tereso del Río[1] and Matthew England[2]

Coventry University

8th SC^2 International Workshop
July 28, 2023 - Tromsø, Norway

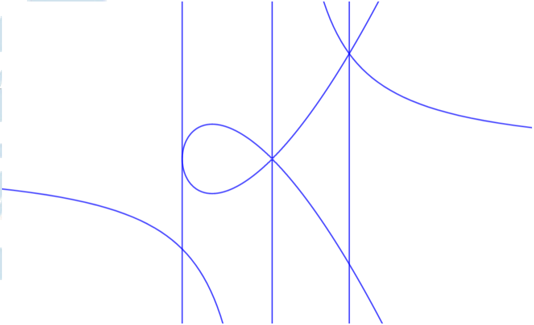
[1] Supported by Coventry University.
[2] Supported by UKRI EPSRC Grant EP/T015748/1, "*Pushing Back the Doubly-Exponential Wall of Cylindrical Algebraic Decomposition*" (DEWCAD).

Outline

Introduction to CAD

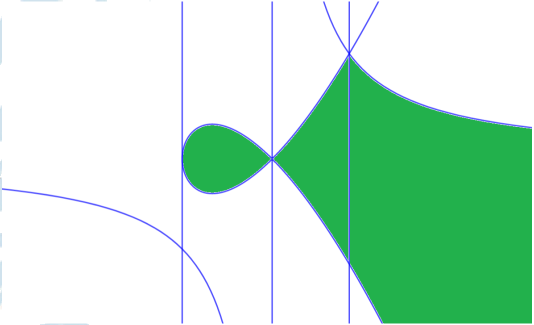
Given a set of polynomials

$$S = \{xy - 1, y^2 - x^3 - x^2\}$$



Introduction to CAD

We may want to know where $xy - 1 < 0$ and $y^2 - x^3 - x^2 < 0$.



The only implemented general-purpose algorithm that guarantees to answer such questions is CAD, firstly proposed in [Collins(1975)].

Pros and cons

- Useful in biology [Röst and Sadeghimanesh(2021)], robotics, proving mathematical inequalities [Gerhold and Kauers(2006)], ...
- Davenport proved in [Davenport and Heintz(1988)] that CAD has doubly exponential complexity with respect to the number of variables.
- and that is SCARY!

Pros and cons

- Useful in biology [Röst and Sadeghimanesh(2021)], robotics, proving mathematical inequalities [Gerhold and Kauers(2006)], ...
- Davenport proved in [Davenport and Heintz(1988)] that CAD has doubly exponential complexity with respect to the number of variables.
- and that is SCARY!

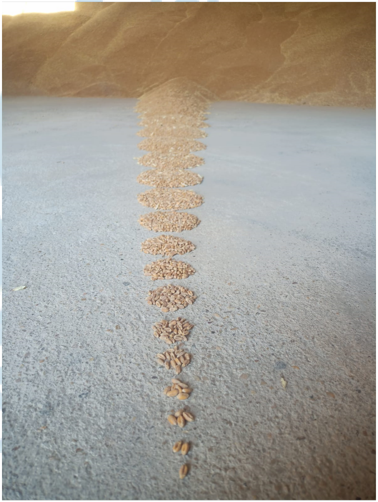
Pros and cons

- Useful in biology [Röst and Sadeghimanesh(2021)], robotics, proving mathematical inequalities [Gerhold and Kauers(2006)], ...
- Davenport proved in [Davenport and Heintz(1988)] that CAD has doubly exponential complexity with respect to the number of variables.
- and that is SCARY!

Pros and cons

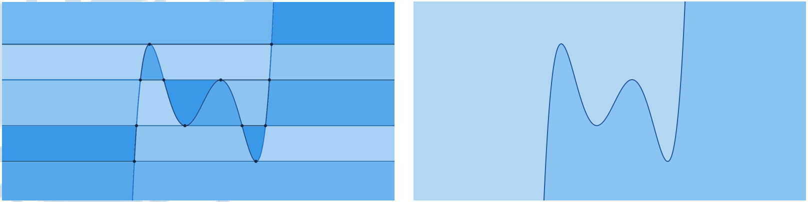
- Useful in biology [Röst and Sadeghimanesh(2021)], robotics, proving mathematical inequalities [Gerhold and Kauers(2006)], ...
- Davenport proved in [Davenport and Heintz(1988)] that CAD has doubly exponential complexity with respect to the number of variables.
- and that is SCARY!

A little story



Variable ordering

Brown and Davenport [Brown and Davenport(2007)]:
Depending on variable ordering, **constant** or **doubly exponential** complexity.



Variable ordering

Choosing the right variable ordering:

- Humans have proposed heuristics for this task: e.g. `sotd` [Dolzmann et al.(2004)Dolzmann, Seidl, and Sturm]; `brown` [Brown(2004)] and `mods` [?]
- Machine Learning models have been trained for this purpose e.g. [?] and [Chen et al.(2020)Chen, Zhu, and Chi]

Variable ordering

Choosing the right variable ordering:

- Humans have proposed heuristics for this task: e.g. `sotd` [Dolzmann et al.(2004)Dolzmann, Seidl, and Sturm]; `brown` [Brown(2004)] and `mods` [?]
- Machine Learning models have been trained for this purpose e.g. [?] and [Chen et al.(2020)Chen, Zhu, and Chi]
.

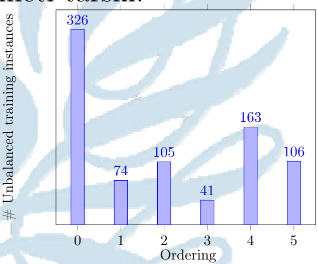
Training models

In [England and Florescu(2019)] multiple models were trained.

Name	**Accuracy**
`brown`	0.553
`gmods`	0.563
KNN	0.555
DT	0.573
SVC	0.549
MLP	0.569

A glance at the dataset

Extracted from QFNRA problems of the SMT-LIB; mainly meti-tarski.



Changing a label

If the optimal ordering for $\{x_1x_2^3 + x_2^2x_3^2, x_2x_3^3 - 1\}$ is 0.

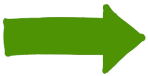
The six possible variable orderings

Ordering Name	**Ordering**
Ordering 0	$x_1 \succ x_2 \succ x_3$
Ordering 1	$x_1 \succ x_3 \succ x_2$
Ordering 2	$x_2 \succ x_1 \succ x_3$
Ordering 3	$x_2 \succ x_3 \succ x_1$
Ordering 4	$x_3 \succ x_1 \succ x_2$
Ordering 5	$x_3 \succ x_2 \succ x_1$

By simply swapping the names of x_1 and x_2 we get an instance with optimal ordering 2: $\{x_2x_1^3 + x_1^2x_3^2, x_1x_3^3 - 1\}$.

Analogy with arrows for computer vision

Normally, we cannot change the labels on demand but our problem is symmetric.

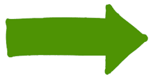


Arrow pointing right

Arrow pointing up

Analogy with arrows for computer vision

Normally, we cannot change the labels on demand but our problem is symmetric.

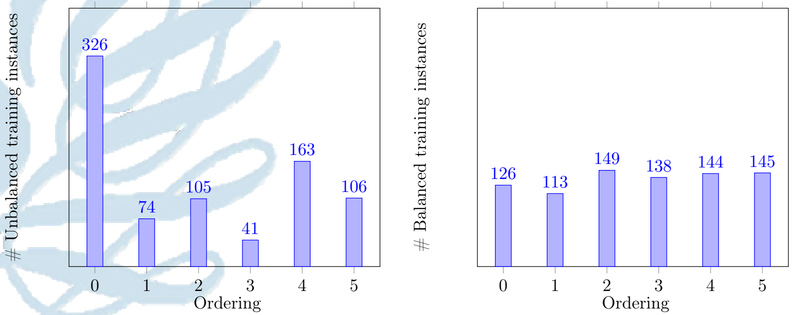


Arrow pointing right

Arrow pointing down

Balancing the dataset

By randomly permuting variables in an instances we can balance our datasets.



Problems caused by unbalancedness

Models trained on unbalanced data do not perform well on balanced data.

Testing dataset	Unbalanced	Balanced
KNN-Unbalanced	0.51	0.21
DT-Unbalanced	0.53	0.31
SVC-Unbalanced	0.48	0.23
RF-Unbalanced	**0.58**	**0.35**
MLP-Unbalanced	0.51	0.32

Accuracy of models trained on the unbalanced dataset, when tested on the different testing datasets.

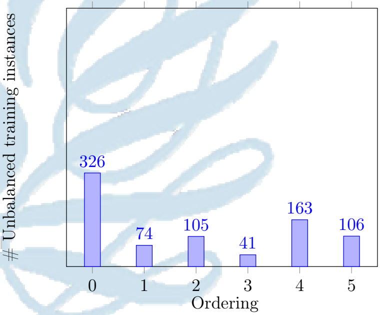
Balancing solves this issue

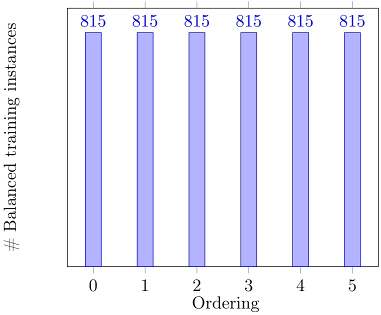
Testing dataset	Unbalanced	Balanced
KNN-Balanced	0.41	0.36
DT-Balanced	0.43	0.45
SVC-Balanced	0.25	0.3
RF-Balanced	**0.49**	**0.52**
MLP-Balanced	0.45	0.43

Accuracy of models trained on the balanced dataset, when tested on the different testing datasets.

Augmenting the dataset

Including all possible permutations we can augmentate the dataset.



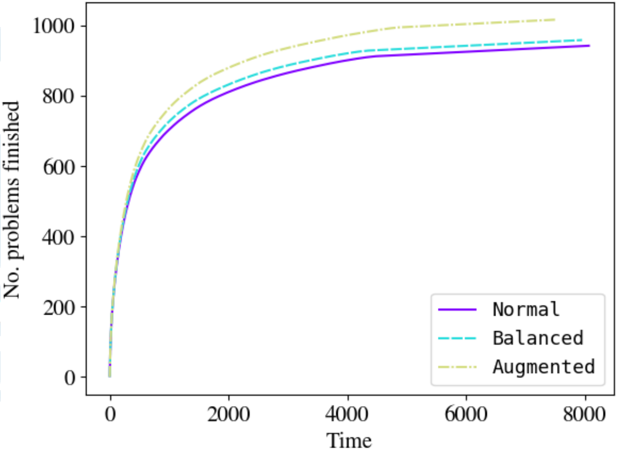


Augmenting boosts the results

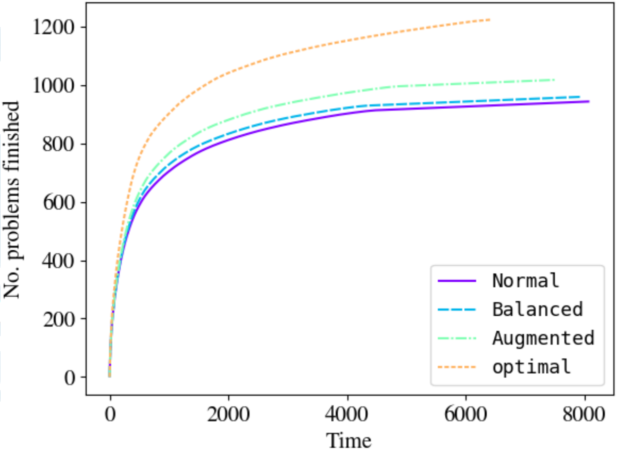
Testing dataset	Unbalanced	Balanced
KNN-Augmented	0.54	0.55
DT-Augmented	0.54	0.55
SVC-Augmented	0.46	0.48
RF-Augmented	**0.62**	**0.63**
MLP-Augmented	0.48	0.5

Accuracy of models trained on the augmented dataset, when tested on the different testing datasets.

Survival plot SVC



Survival plot SVC



Comparing accuracies

Training dataset	Normal	Balanced	Augmented
KNN	0.3	0.42	**0.55**
DT	0.35	0.43	**0.54**
MLP	0.35	0.45	**0.47**
SVC	0.23	0.29	**0.48**
RF	0.46	0.53	**0.61**

Accuracy of models on the balanced testing dataset, having been trained on the different training datasets.

Comparing timings

Training dataset	Normal	Balanced	Augmented
KNN	21 603	20 927	18 850
DT	20 352	17 299	17 404
SVC	25 004	23 913	19 980
RF	19 909	17 391	16 301
MLP	21 977	20 210	18 509

Accuracy of models on the balanced testing dataset, having been trained on the different training datasets.

Comparison with [Hester et al.(2023)Hester, Hitaj, Passmore, Owre, Shankar

- Very similar results.
- I removed loads of repeated examples (around 8000 vs around 1000).
- I used some more features.
- I still have to check if those two make any difference.

Future work

- Extra augmentation methods.
- Using regression instead of classification.
- Using reinforcement learning (pick one variable at a time).
- Encode sets of polynomials as graph and using Graph NN.

Future work

- Extra augmentation methods.
- Using regression instead of classification.
- Using reinforcement learning (pick one variable at a time).
- Encode sets of polynomials as graph and using Graph NN.

Future work

- Extra augmentation methods.
- Using regression instead of classification.
- Using reinforcement learning (pick one variable at a time).
- Encode sets of polynomials as graph and using Graph NN.

Future work

- Extra augmentation methods.
- Using regression instead of classification.
- Using reinforcement learning (pick one variable at a time).
- Encode sets of polynomials as graph and using Graph NN.

Comparing with regression

Classification	
Name	**Time**
KNN	18 850
DT	17 404
SVC	19 980
RF	16 301
MLP	18 509

Regression	
Name	**Time**
DTR	17 206
SVR	26 100
RFR	11 391
KNNR	15 362
MLPR	25 219

Timings for different paradigms

Lesson to take from this talk

Representations of mathematical objects often have symmetries and those can be exploited to augmentate the number of representations that we have of a given object.
Very rarely we can give a mathematical object to a machine learning model (variable length), and augmentation is a tool to give as many views of the same object as possible.

Bibliography I

Christopher W Brown.
Companion to the Tutorial Cylindrical Algebraic Decomposition.
In *International Symposium on Symbolic and Algebraic Computation - ISSAC*, 2004.

Christopher W. Brown and James H. Davenport.
The complexity of quantifier elimination and cylindrical algebraic decomposition.
Proceedings of the International Symposium on Symbolic and Algebraic Computation, ISSAC, pages 54–60, 2007.
doi: 10.1145/1277548.1277557.
Publisher: ACM Press ISBN: 9781595937438.

Bibliography II

Changbo Chen, Zhangpeng Zhu, and Haoyu Chi.
Variable Ordering Selection for Cylindrical Algebraic Decomposition with Artificial Neural Networks.
In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12097 LNCS, pages 281–291. Springer, 2020.
ISBN 978-3-030-52199-8.
doi: 10.1007/978-3-030-52200-1_28.
ISSN: 16113349.

Bibliography III

George E. Collins.
Quantifier elimination for real closed fields by cylindrical algebraic decomposition.
Lecture Notes in Computer Science, 33(Automata Theory and Formal Languages):134–183, 1975.
ISSN 16113349.
doi: 10.1007/3-540-07407-4_17.
Publisher: Springer Verlag ISBN: 9783540074076.

Bibliography IV

- James H. Davenport and Joos Heintz.
  Real quantifier elimination is doubly exponential.
  Journal of Symbolic Computation, 5(1-2):29–35, February 1988.
  ISSN 07477171.
  doi: 10.1016/S0747-7171(88)80004-X.
  Publisher: Academic Press.
- Andreas Dolzmann, Andreas Seidl, and Thomas Sturm.
  Efficient projection orders for CAD.
  In *Proceedings of the 2004 International Symposium on Symbolic and Algebraic Computation - ISSAC*, pages 111–118, New York, New York, USA, 2004. ACM Press.
  ISBN 1-58113-827-X.
  doi: 10.1145/1005285.1005303.

Bibliography V

Matthew England and Dorian Florescu.
Comparing Machine Learning Models to Choose the Variable Ordering for Cylindrical Algebraic Decomposition.
In Cezary Kaliszyk, Edwin Brady, Andrea Kohlhase, and Claudio Sacerdoti Coen, editors, *Intelligent Computer Mathematics*, volume 11617 of *Lecture Notes in Computer Science*, pages 93–108. Springer International Publishing, Cham, 2019.
ISBN 978-3-030-23249-8 978-3-030-23250-4.
doi: 10.1007/978-3-030-23250-4_7.
URL `http://link.springer.com/10.1007/978-3-030-23250-4_7`.

Bibliography VI

- Stefan Gerhold and Manuel Kauers.
  A computer proof of Turán's inequality.
  Journal of Inequalities in Pure and Applied Mathematics, 2006.

- John Hester, Briland Hitaj, Grant Passmore, Sam Owre, Natarajan Shankar, and Eric Yeh.
  Revisiting Variable Ordering for Real Quantifier Elimination using Machine Learning.
  arXiv preprint, 2023.
  doi: 10.48550/ARXIV.2302.14038.
  URL https://arxiv.org/abs/2302.14038.
  Publisher: arXiv Version Number: 1.

Bibliography VII

Gergely Röst and Amirhosein Sadeghimanesh.
Exotic Bifurcations in Three Connected Populations with Allee Effect.
International Journal of Bifurcation and Chaos, 31(13), 2021.
ISSN 02181274.
doi: 10.1142/S0218127421502023.
Publisher: World Scientific Publishing Company.

Apendix