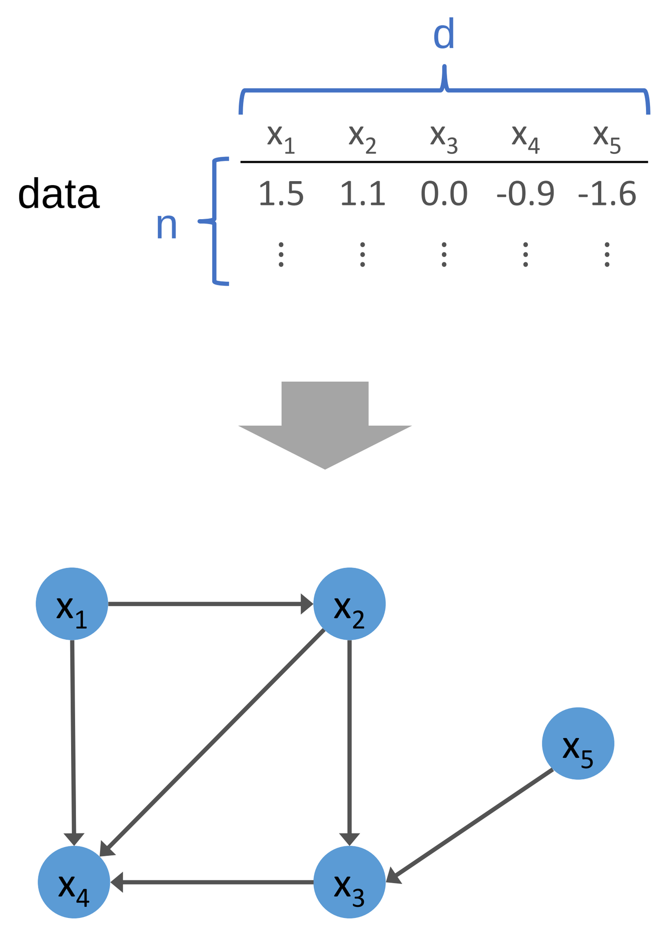
DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks

Presented by: **Yue Yu**, Department of Mathematics, Lehigh University

Joint work with: **Dennis Wei, Tian Gao**, IBM research

Bayesian Networks via Continuous Optimization



Given n samples of $X \in \mathbb{R}^d$, learn Bayesian network parametrized by W

Score-based learning:

$$\begin{array}{ll} \text{minimize} & \text{score } F(W) \\ \text{subject to} & G(W) \in \text{DAGs} \end{array}$$

combinatorial

NOTEARS [1]:

$$\begin{array}{ll} \text{minimize} & \text{score } F(W) \\ \text{subject to} & h(A) = 0 \end{array}$$

continuous, differentiable

adjacency matrix

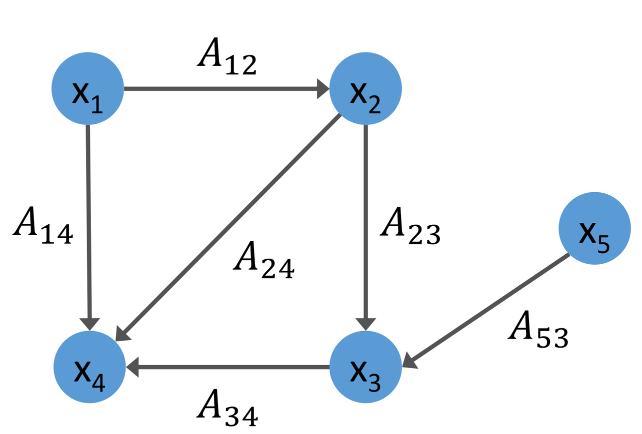
Our Contributions

Theoretical understanding of continuous optimization framework, leading to significant algorithmic improvements

1. Better understanding of NOTEARS
2. Understanding of KKT optimality conditions for reformulation
3. KKT-search post-processing improves all tested algorithms

General Class of Acyclicity Constraints

Adjacency matrix A:
$A_{ij} > 0 \Leftrightarrow$ edge



NOTEARS [1]: $h(A) = \mathrm{tr}(e^A) - d$

DAG-GNN [2]: $h(A) = \mathrm{tr}\left((I + A/d)^d\right) - d$

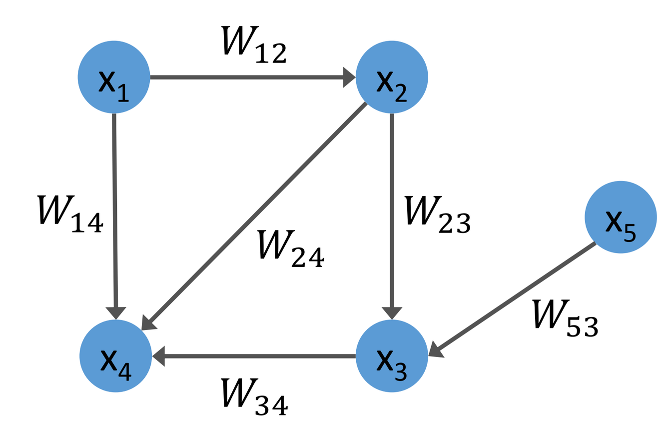
Our generalization:

$$h(A) = \mathrm{tr}\left(\sum_{p=1}^{d} c_p A^p\right) \qquad c_p > 0$$

$\nabla h(A)$ has a directed walk interpretation

1. Better Understanding of NOTEARS

Assumption: Each edge corresponds to one parameter W_{ij}
e.g. generalized linear SEM

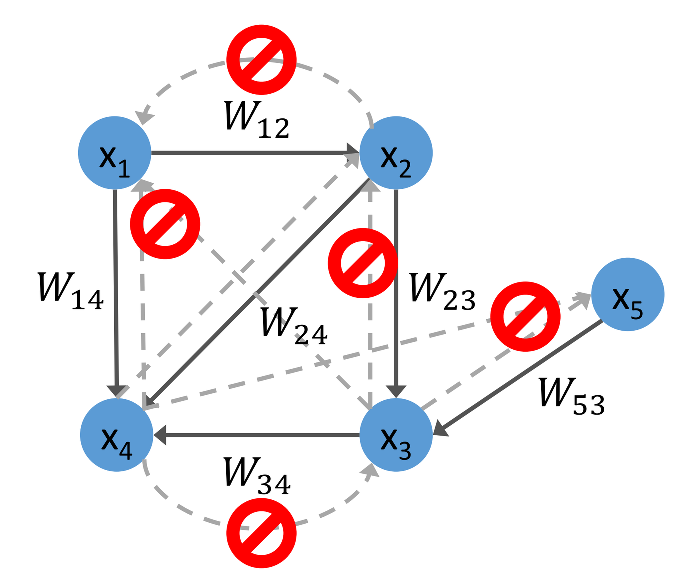


- Acyclic solutions cannot satisfy KKT conditions
- Augmented Lagrangian algorithm cannot converge to acyclic solution even with high penalty

$$\begin{array}{ll} \text{minimize} & F(W) \\ \text{subject to} & h(\underbrace{W \circ W}) = 0 \end{array}$$

quadratic adjacency matrix $A = W \circ W$

2. KKT Conditions for Reformulation



- KKT conditions are indeed necessary: Local minima must satisfy them
- If score $F(W)$ is convex, KKT conditions **sufficient** for local minimality, despite non-convexity of constraint
- Understanding of KKT conditions in terms of edge absence constraints: Collectively sufficient, individually necessary in preventing cycles

$$\begin{array}{ll} \text{minimize} & F(W) \\ \text{subject to} & h(\underbrace{|W|}) = 0 \end{array}$$

absolute adjacency matrix $A = |W|$

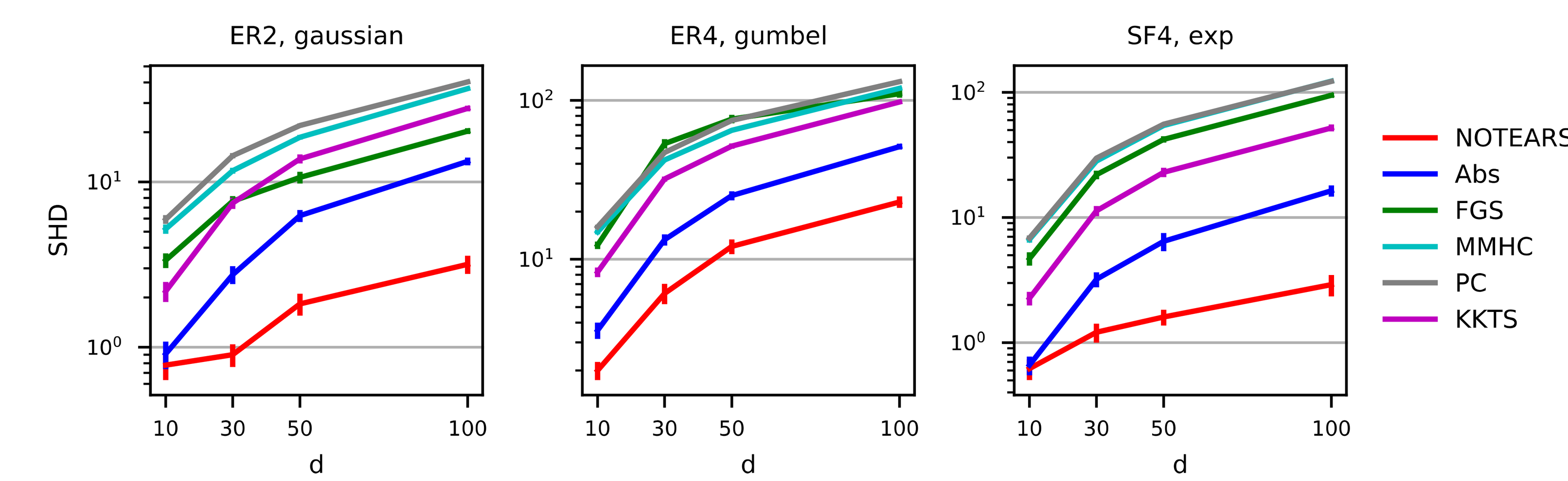
$$\begin{array}{ll} \text{minimize} & F(W) \\ \text{subject to} & W_{ij} = 0, \quad (i,j) \in \mathcal{Z} \end{array}$$

Proposed Algorithms

1. Augmented Lagrangian with $A = |W|$ ('Abs')
2. KKT-search to satisfy KKT conditions

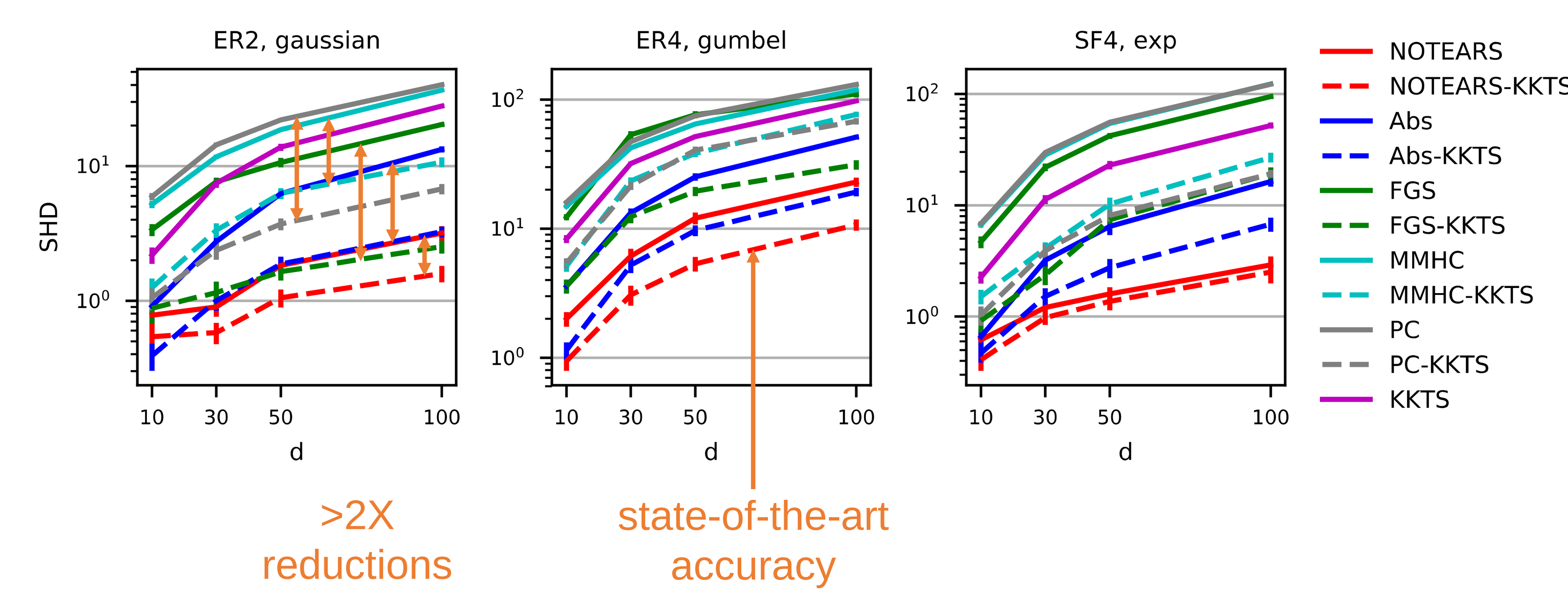
3. KKT-search Improves Existing Algorithms

Base algorithms: NOTEARS still best, Abs second



Structural Hamming distance (SHD) with respect to true graph for different graph types and $n = 1000$ samples

With KKT-search:
Consistent improvement across base algorithms and dimension d



References

[1] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, Eric P. Xing (2018). "DAGs with NO TEARS: Continuous optimization for structure learning." In Advances in Neural Information Processing Systems (NeurIPS).

[2] Yue Yu, Jie Chen, Tian Gao, Mo Yu (2019). "DAG-GNN: DAG structure learning with graph neural networks.". In International Conference on Machine Learning (ICML).

[3] Sébastian Lachapelle, Philippe Brouillard, Tristan Deleu, Simon Lacoste-Julien (2020). "Gradient-based neural DAG learning". In International Conference on Learning Representations (ICLR).

[4] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, Eric P. Xing (2020). "Learning sparse nonparametric DAGs." In International Conference on Artificial Intelligence and Statistics (AISTATS).

For the full paper please see: Dennis Wei, Tian Gao, and Yue Yu. "DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks." *Advances in Neural Information Processing Systems* 33 (2020).