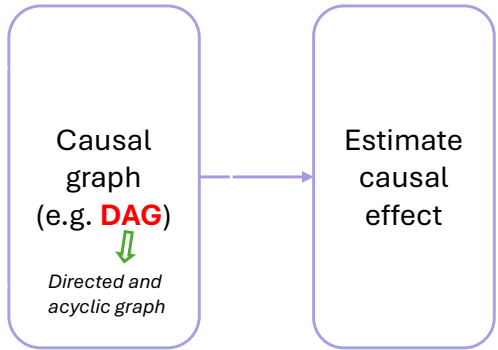
A Convex Mixed-Integer Programming Framework for Causal Additive Models

Xiaozhu Zhang

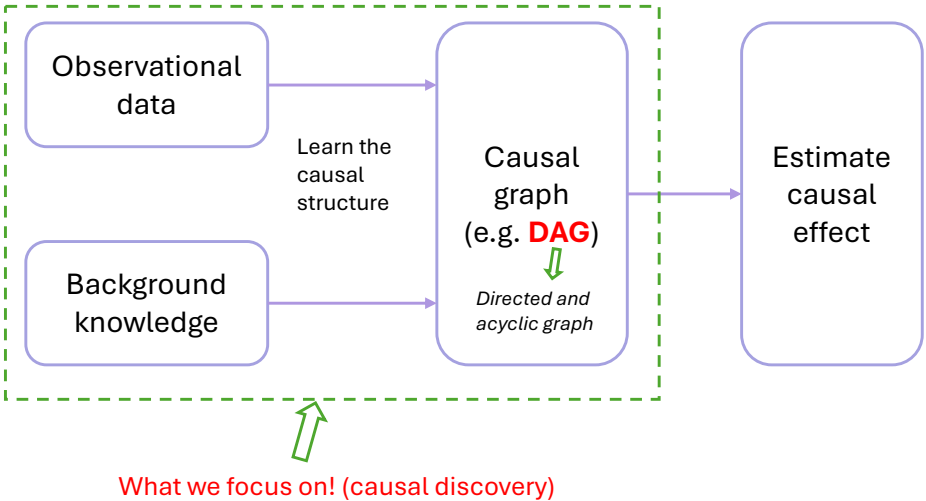
Department of Statistics, University of Washington

Joint with Nir Keret, Ali Shojaie, Rahul Mazumder, Armeen Taeb

Typical causal inference pipeline



Typical causal inference pipeline



Convexity and discrete model selection

- Convexity is the cornerstone of computation in statistics
  - For some discrete model selection problems,
    - ✔ convex relaxations (e.g. Lasso)
  - For some other problems (like our problem in this talk),
    - ✘ convex relaxations (Evans, 2020)

Convexity and discrete model selection

- Convexity is the cornerstone of computation in statistics
  - For some discrete model selection problems,
    - ✔ convex relaxations (e.g. Lasso)
  - For some other problems (like our problem in this talk),
    - ✘ convex relaxations (Evans, 2020)

- Convex mixed integer programming (MIP)
  - discrete structure $\Leftrightarrow$ binary variable
  - when binary variable is relaxed to $[0, 1]$ $\Leftrightarrow$ convex

Convexity and discrete model selection

- Convexity is the cornerstone of computation in statistics
  - For some discrete model selection problems,
    - ✔ convex relaxations (e.g. Lasso)
  - For some other problems (like our problem in this talk),
    - ✘ convex relaxations (Evans, 2020)

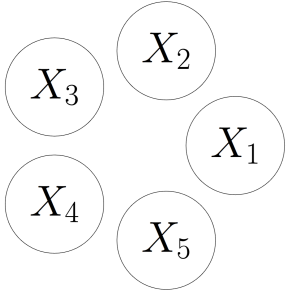
- Convex mixed integer programming (MIP)
  - discrete structure $\Leftrightarrow$ binary variable
  - when binary variable is relaxed to $[0, 1]$ $\Leftrightarrow$ convex

- Why is the "relaxed" convexity important in MIP?
  - certify optimality
  - prune search space

Goal: non-linear DAG discovery

Given data matrix $X \in \mathbb{R}^{n \times p}$ (n obs. and p nodes)

How to learn the relationship between the p nodes?

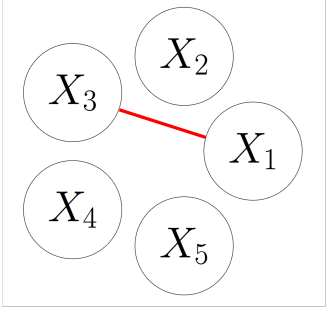


Goal: non-linear DAG discovery

Given data matrix $X \in \mathbb{R}^{n \times p}$ (n obs. and p nodes)

How to learn the relationship between the p nodes?

- **Presence** of direct causal effect?
  $k \overset{?}{—} j$

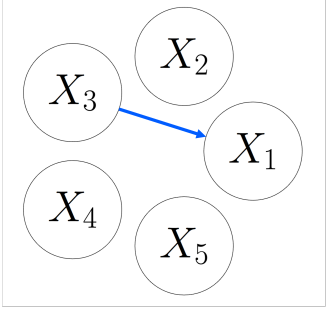


Goal: non-linear DAG discovery

Given data matrix $X \in \mathbb{R}^{n \times p}$ (n obs. and p nodes)

How to learn the relationship between the p nodes?

- **Presence** of direct causal effect?
  $k \overset{?}{—} j$
- If yes, what's the **direction**?
  $k \to j$ or $j \to k$?

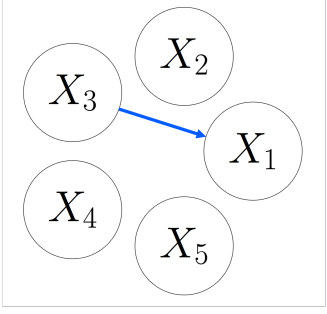


Goal: non-linear DAG discovery

Given data matrix $X \in \mathbb{R}^{n \times p}$ (n obs. and p nodes)

How to learn the relationship between the p nodes?

- **Presence** of direct causal effect?
  $k \overset{?}{-} j$
- If yes, what's the **direction**?
  $k \to j$ or $j \to k$?



This talk: causal relationships are **non-linear**!

Goal: a flexible framework with guarantees

Weaknesses of existing methods:

	Optimality guarantee	Statistical guarantee	Allows heteroskedasticity
CAM (Bühlmann et al., 2014)	✗	✔	
NoTears (Zhang et al., 2018)	✗	✔	
NPVAR (Gao et al., 2020)			✗

Goal: a flexible framework with guarantees

Weaknesses of existing methods:

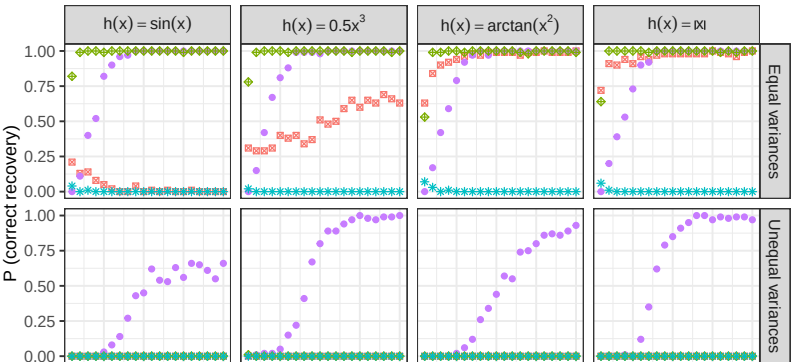
	Optimality guarantee	Statistical guarantee	Allows heteroskedasticity
CAM (Bühlmann et al., 2014)	✗	✔	
NoTears (Zhang et al., 2018)	✗	✔	
NPVAR (Gao et al., 2020)			✗

Our framework:

- with **optimality** guarantees
- with **statistical** guarantees
- **Allows** for heteroskedasticity

Preview

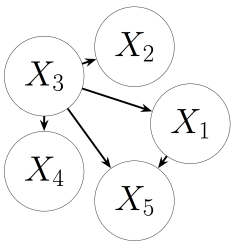
For a 5-variable toy example:



Takeaway:

- Under equal-variance, only NPVAR and our method recover the true DAG
- Under unequal-variance, only our method recovers the true DAG w.h.p.

Model: additive SEM with Gaussian errors



$\Longleftrightarrow$

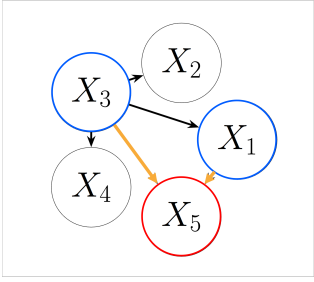
The SEM. For $j = 1, \ldots, p$,

$$X_j = \sum_{k \in \mathrm{pa}(j)} f_{kj}^\star(X_k) + \epsilon_j,$$

$$\epsilon_j \sim \mathcal{N}(0, \sigma_j^{\star 2}), \quad \sigma_j^\star > 0,$$

$f_{kj}^\star$ is non-linear, $\quad \mathbb{E}[f_{kj}^\star(X_k)] = 0, \ \forall j, k$

Model: additive SEM with Gaussian errors



$\Longleftrightarrow$

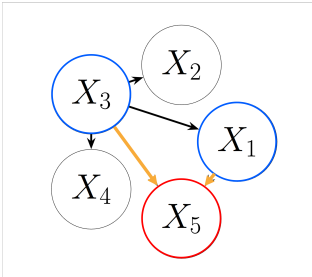
The SEM. For $j = 1, \ldots, p$,

$$X_j = \sum_{k \in \mathrm{pa}(j)} f_{kj}^{\star}(X_k) + \epsilon_j,$$

$$\epsilon_j \sim \mathcal{N}(0, \sigma_j^{\star 2}), \quad \sigma_j^{\star} > 0,$$

$f_{kj}^{\star}$ is non-linear, $\quad \mathbb{E}[f_{kj}^{\star}(X_k)] = 0, \ \forall j, k$

Model: additive SEM with Gaussian errors



$\Longleftrightarrow$

The SEM. For $j = 1, \ldots, p$,

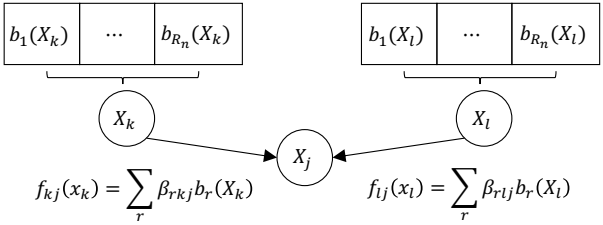
$$X_j = \sum_{k \in \mathrm{pa}(j)} f_{kj}^{\star}(X_k) + \epsilon_j,$$

$$\epsilon_j \sim \mathcal{N}(0, \sigma_j^{\star 2}), \quad \sigma_j^{\star} > 0,$$

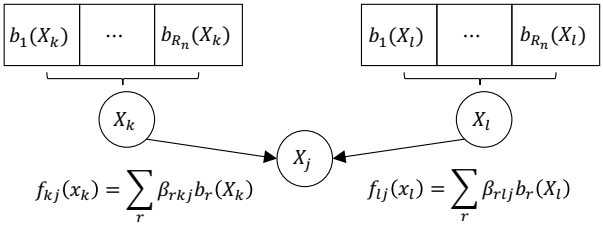
$f_{kj}^{\star}$ is non-linear, $\quad \mathbb{E}[f_{kj}^{\star}(X_k)] = 0, \ \forall j, k$

- **Faithfulness**: $f_{kj}^{\star} \not\equiv 0$ if and only if $k \in \mathrm{pa}(j)$
- **Identifiability** of DAG: when $f_{kj}^{\star}$ are non-linear and in C^3 (Peters et al., 2014)

Approaching the non-linearity: basis functions



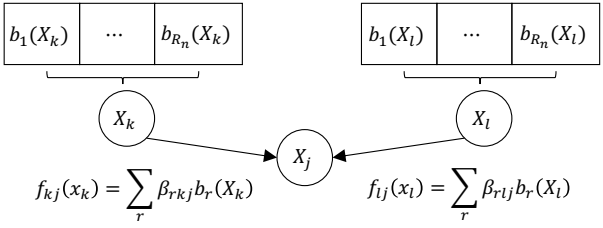
Approaching the non-linearity: basis functions



$$Z_j = \begin{bmatrix} b_1(X_1) & \cdots & b_1(X_p) & \cdots & b_{R_n}(X_1) & \cdots & b_{R_n}(X_p) \end{bmatrix} \in \mathbb{R}^{n \times pR_n}$$

$$\beta_j = \begin{bmatrix} \beta_{11j} & \cdots & \beta_{1pj} & \cdots & \beta_{R_n 1j} & \cdots & \beta_{R_n pj} \end{bmatrix} \in \mathbb{R}^{pR_n}$$

Approaching the non-linearity: basis functions



- Group sparsity:
  If $k \not\to j$,
  then $\beta_{rkj} = 0, \ \forall r$
- $\sum_k f_{kj}(x_k) = Z_j \beta_j$

$$Z_j = \begin{bmatrix} b_1(X_1) & \cdots & b_1(X_p) & \cdots & b_{R_n}(X_1) & \cdots & b_{R_n}(X_p) \end{bmatrix} \in \mathbb{R}^{n \times pR_n}$$

$$\beta_j = \begin{bmatrix} \beta_{11j} & \cdots & \beta_{1pj} & \cdots & \beta_{R_n 1j} & \cdots & \beta_{R_n pj} \end{bmatrix} \in \mathbb{R}^{pR_n}$$

Approaching the non-linearity: basis functions

Negative log-likelihood.

$$\ell_n(\beta, \sigma) = \sum_{j=1}^{p} \log \sigma_j^2 + \sum_{j=1}^{p} \frac{\|X_j - Z_j \beta_j\|_n^2}{\sigma_j^2}$$

Can we directly minimize it over (β, σ)?

No! b/c acyclicity and sparsity!

- Group sparsity:
  If $k \not\to j$,
  then $\beta_{rkj} = 0, \ \forall r$
- $\sum_k f_{kj}(x_k) = Z_j \beta_j$

Approaching the non-linearity: basis functions

Negative log-likelihood.

$$\ell_n(\beta, \sigma) = \sum_{j=1}^{p} \log \sigma_j^2 + \sum_{j=1}^{p} \frac{\|X_j - Z_j \beta_j\|_n^2}{\sigma_j^2}$$

Can we directly minimize it over (β, σ)?
No! b/c acyclicity and sparsity!

- Group sparsity:
  If $k \not\to j$,
  then $\beta_{rkj} = 0, \ \forall r$
- $\sum_k f_{kj}(x_k) = Z_j \beta_j$

A mixed-integer programming (MIP) formulation

- **Edge selection**:

  Edge presence $g_{kj} \in \{0, 1\}$

  $k \not\to j \iff g_{kj} = 0 \iff \beta_{rkj} = 0, \ \forall r$

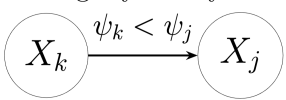
A mixed-integer programming (MIP) formulation

- **Edge selection**:
  Edge presence $g_{kj} \in \{0, 1\}$
  $k \not\to j \iff g_{kj} = 0 \iff \beta_{rkj} = 0, \ \forall r$

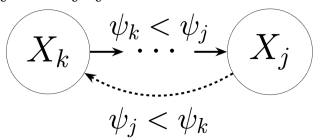
- **Acyclicity**:
  Layer value $\psi_j \in [1, p]$

$$1 - p + p g_{kj} \leq \psi_j - \psi_k$$

direct edges yield layer order



cyclicity yields contradiction



A mixed-integer programming (MIP) formulation

- **Edge selection**:
  Edge presence $g_{kj} \in \{0, 1\}$
  $k \not\to j \iff g_{kj} = 0 \iff \beta_{rkj} = 0, \ \forall r$

- **Acyclicity**:
  Layer value $\psi_j \in [1, p]$

$$1 - p + p g_{kj} \leq \psi_j - \psi_k$$

- **Regularization**:
  $\lambda_n \sum_{kj} g_{kj}$ (DAG sparsity), ridge penalty on β

The MIP formulation. g: binary β: continuous ψ: layer value

$$\min_{\beta,g,\psi} \underbrace{\sum_{j=1}^{p} \log\left[\|X_j - Z_j\beta_j\|_n^2 + \gamma_n\|\beta_j\|_2^2\right]}_{F(\beta,g)} + \lambda_n \underbrace{\sum_{k,j} g_{kj}}_{\text{DAG sparsity}},$$

$$\text{s.t.} \quad (1 - g_{kj})\beta_{rkj} = 0, \quad \forall r, k, j \quad \text{(edge selection)}$$

$$1 - p + pg_{kj} \leq \psi_j - \psi_k, \quad \forall k, j \quad \text{(acyclicity)}$$

From background knowledge (optional):

$g_{ab} = 1$ (if we know $a \to b$)

$\psi_c < \psi_d$ (if we know c is before d)

The MIP formulation. g: binary β: continuous ψ: layer value

$$\min_{\beta,g,\psi} \underbrace{\sum_{j=1}^{p} \log\left[\|X_j - Z_j\beta_j\|_n^2 + \gamma_n\|\beta_j\|_2^2\right]}_{F(\beta,g)} + \lambda_n \underbrace{\sum_{k,j} g_{kj}}_{\text{DAG sparsity}},$$

$$\text{s.t.} \quad (1 - g_{kj})\beta_{rkj} = 0, \quad \forall r, k, j \quad \text{(edge selection)}$$

$$1 - p + pg_{kj} \leq \psi_j - \psi_k, \quad \forall k, j \quad \text{(acyclicity)}$$

From **background knowledge** (optional) :

$g_{ab} = 1$ (if we know $a \to b$)

$\psi_c < \psi_d$ (if we know c is an ancestor of d)

Challenges and opportunities of MIP

- MIP is NP-hard in general

Challenges and opportunities of MIP

- MIP is NP-hard in general
- The good news is,

  good exploration of the **problem structure** $\Longrightarrow$ scalable algorithm

  - ℓ_0-regression (Bertsimas & Van Parys, 2020)
  - multi-task learning (Behdin et al., 2025)
  - differential privacy (Prastakos et al., 2025)

Challenges and opportunities of MIP

- MIP is NP-hard in general
- The good news is,

  good exploration of the **problem structure** $\Longrightarrow$ scalable algorithm

  - ℓ_0-regression (Bertsimas & Van Parys, 2020)
  - multi-task learning (Behdin et al., 2025)
  - differential privacy (Prastakos et al., 2025)
- What structures do these settings explore?
  - Convexity (after relaxing binary variables)
  - Many non-binary variables can be projected out (and retain convexity)

Our setting is very different... How can we identify a useful structure?

Structure in our problem

$$\min_{\beta,g,\psi} F(\beta,g) + \lambda \sum_{k,j} g_{kj} \quad \text{s.t. edge selection } (\beta,\, g) + \text{acyclicity constrains } (g,\, \psi)$$

- We project out the continuous β in $F(\beta, g)$:

$$f(g) = \min_{\beta} F(\beta, g) \;\; \text{s.t. edge selection constraint}$$

Structure in our problem

$$\min_{\beta,g,\psi} F(\beta,g) + \lambda \sum_{k,j} g_{kj} \quad \text{s.t. edge selection } (\beta,\, g) + \text{acyclicity constrains } (g,\, \psi)$$

- We project out the continuous β in $F(\beta, g)$:

$$f(g) = \min_{\beta} F(\beta, g) \quad \text{s.t. edge selection constraint}$$

When fixing the DAG structure via g:

$$F(\beta, g) = \sum_{j=1}^{p} \log \left[\|X_j - Z_j \beta_j\|_n^2 + \gamma \|\beta_j\|_2^2 \right]$$

p independent ridge regression $\implies$ This can be solved in **closed-form**!

$$f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j), \qquad A(g_j) \succ 0 \text{ and linear in } g_j$$

Structure in our problem

$$\min_{g,\psi} f(g) + \lambda \sum_{k,j} g_{kj} \qquad \text{s.t. acyclicity constrains } (g,\ \psi)$$

- We project out the continuous β in $F(\beta, g)$:

$$f(g) = \min_{\beta} F(\beta, g) \quad \text{s.t. edge selection constraint}$$

When fixing the DAG structure via g:

$$F(\beta, g) = \sum_{j=1}^{p} \log \left[\|X_j - Z_j \beta_j\|_n^2 + \gamma \|\beta_j\|_2^2 \right]$$

p independent ridge regression $\implies$ This can be solved in **closed-form**!

$$f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j), \qquad A(g_j) \succ 0 \text{ and linear in } g_j$$

Structure in our problem

$$\min_{g,\psi} f(g) + \lambda \sum_{k,j} g_{kj} \qquad \text{s.t. acyclicity constrains } (g, \psi)$$

- We prove that $f(g)$ is **convex** in g

Structure in our problem

$$\min_{g,\psi} f(g) + \lambda \sum_{k,j} g_{kj} \qquad \text{s.t. acyclicity constrains } (g, \psi)$$

- We prove that $f(g)$ is **convex** in g

Sketch: After some algebra, for any $\nu_j > 0$,

$$f(g) \geq \sum_{j=1}^{p} \left[1 + \log \nu_j - \frac{\overbrace{\nu_j}^{h_j(g)}}{X_j^\top A^{-1}(g_j) X_j} \right] = \sum_{j=1}^{p} \left[1 + \log \nu_j - \nu_j \min_{z: X_j^\top z = 1} z^\top A(g_j) z \right]$$

Structure in our problem

$$\min_{g,\psi} f(g) + \lambda \sum_{k,j} g_{kj} \qquad \text{s.t. acyclicity constrains } (g,\ \psi)$$

- We prove that $f(g)$ is **convex** in g

Sketch: After some algebra, for any $\nu_j > 0$,

$$f(g) \geq \sum_{j=1}^{p} \left[1 + \log \nu_j - \frac{\overbrace{\nu_j}^{h_j(g)}}{X_j^\top A^{-1}(g_j) X_j} \right] = \sum_{j=1}^{p} \left[1 + \log \nu_j - \nu_j \min_{z: X_j^\top z = 1} z^\top A(g_j) z \right]$$

$h_j(g)$ is convex!

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g)\right] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ)\right]$$

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g)\right] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ)\right]$$

Take $\nu_j = X_j^\top A(g_j^\circ)^{-1} X_j$,

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} [1 + \log \nu_j + h_j(g)] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ) \right]$$

Take $\nu_j = X_j^\top A(g_j^\circ)^{-1} X_j$,

- $h_j(g_j^\circ) = -1$

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} [1 + \log \nu_j + h_j(g)] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ) \right]$$

Take $\nu_j = X_j^\top A(g_j^\circ)^{-1} X_j$,

- $h_j(g_j^\circ) = -1$
- $\sum_j \log \nu_j = f(g^\circ)$

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} [1 + \log \nu_j + h_j(g)] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ) \right]$$

Take $\nu_j = X_j^\top A(g_j^\circ)^{-1} X_j$,

- $h_j(g_j^\circ) = -1$
- $\sum_j \log \nu_j = f(g^\circ)$
- $\nabla_{g_j} h_j(g_j)\big|_{g=g^\circ} = \nabla_{g_j} f(g)\big|_{g=g^\circ}$

Now we have:

- $f(g) = \sum_{j=1}^{p} \log(X_j^\top A(g_j)^{-1} X_j)$
- $h_j(g) = -\frac{\nu_j}{X_j^\top A(g_j)^{-1} X_j}$ and it is convex

Then for any fixed g°,

$$f(g) \geq \sum_{j=1}^{p} [1 + \log \nu_j + h_j(g)] \geq \sum_{j=1}^{p} \left[1 + \log \nu_j + h_j(g_j^\circ) + \nabla h_j(g_j^\circ)^\top (g_j - g_j^\circ) \right]$$

Take $\nu_j = X_j^\top A(g_j^\circ)^{-1} X_j$,

- $h_j(g_j^\circ) = -1$
- $\sum_j \log \nu_j = f(g^\circ)$
- $\nabla_{g_j} h_j(g_j)\big|_{g=g^\circ} = \nabla_{g_j} f(g)\big|_{g=g^\circ}$

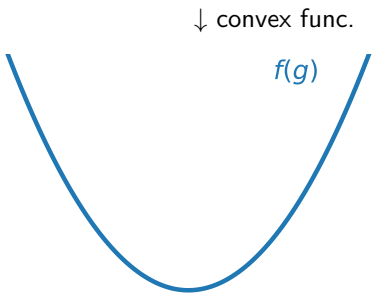
$$f(g) \geq f(g^\circ) + \nabla f(g^\circ)^\top (g - g^\circ)$$

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



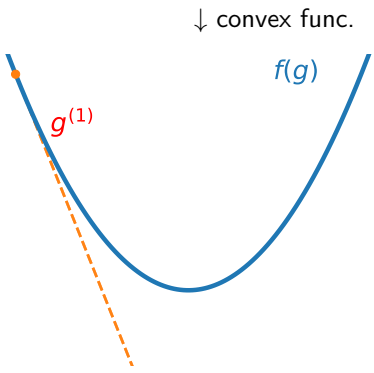
Upper bound u: $+\infty$

Lower bound l: $-\infty$

g can be fractional for the first several rounds

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



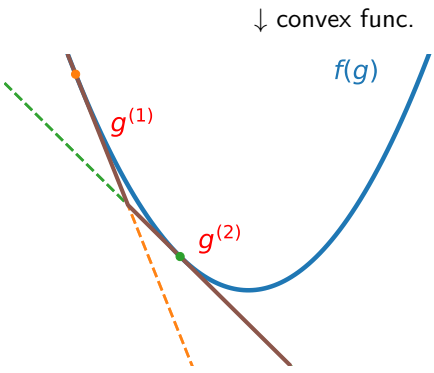
Upper bound u: $f(g^{(1)})$

Lower bound l: $f(1)$

g can be fractional for the first several rounds

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



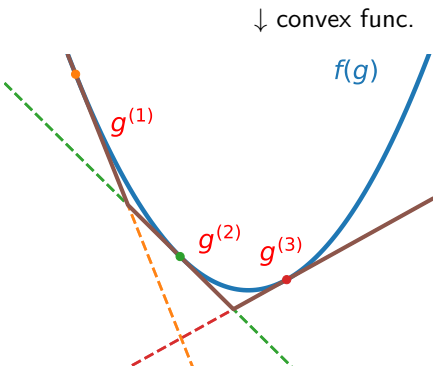
Upper bound u: $f(g^{(2)})$

Lower bound l: $f(1)$

g can be fractional for the first several rounds

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



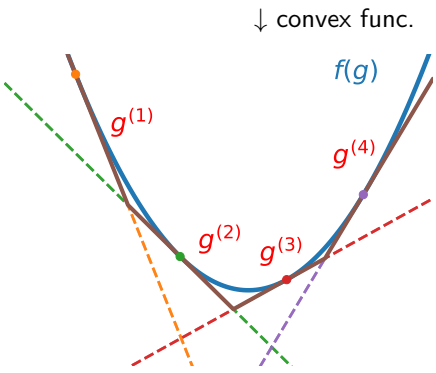
Upper bound u: $f(g^{(3)})$

Lower bound l: $f(0.5)$

g can be fractional for the first several rounds

Cutting plane method

Solve $\min_{g,\psi} f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



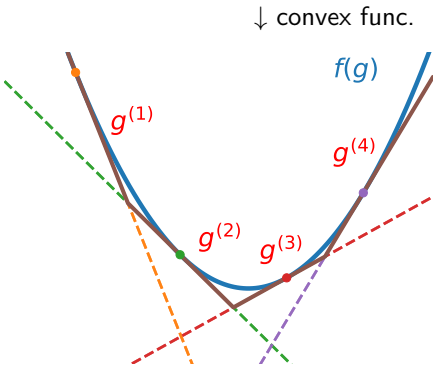
Upper bound u: $f(g^{(3)})$

Lower bound l: $f(0.5)$

g can be fractional for the first several rounds

Cutting plane method

Solve $\min_{g,\psi}$ $f(g) + \lambda \sum_{kj} g_{kj}$ s.t. acyclicity constraint



Upper bound u: $f(g^{(3)})$

Lower bound l: $f(0.5)$

g can be fractional for the first several rounds

A linear MIP in g (scalable)

$$\min_{\theta,g,\psi} \quad \theta + \lambda_n \sum_{k,j} g_{kj}$$

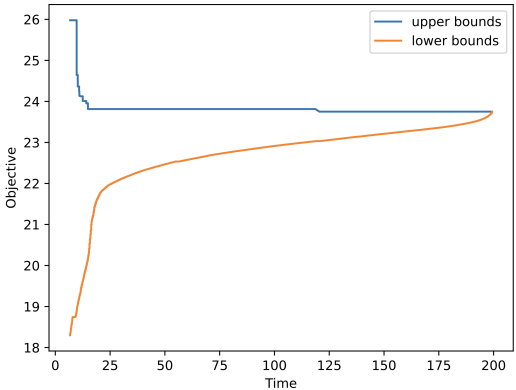
$$\text{s.t.} \quad 1 - p + p g_{kj} \le \psi_j - \psi_k \,, \quad \forall k, j,$$

$$\theta \ge f(g^{(i)}) + \nabla f(g^{(i)})^\top (g - g^{(i)}) \,,$$

$\forall$ cuts i

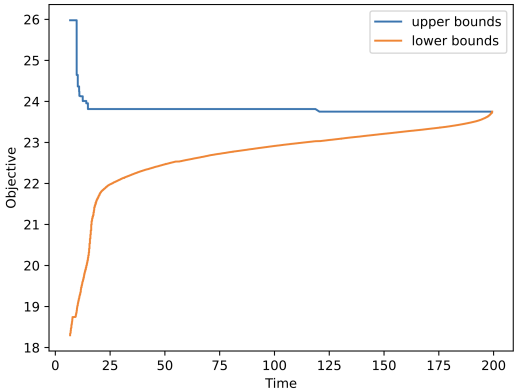
Early stopping

- Increasing lower bound l + decreasing upper bound u
- When $l = u$, optimal solution $\boxed{g^{\text{opt}}}$



Early stopping

- Increasing lower bound l + decreasing upper bound u
- When $l = u$, optimal solution $\boxed{g^{\text{opt}}}$



However, the convergence might be slow

- Upper bound u quickly approaches g^{opt}
- But lower bound l improves slowly
- **Early stop** when $|u - l| \leq \tau^{\text{early}}$
- Early stopping solution $\boxed{g^{\text{early}}}$

Statistical guarantees

Theorem. For sufficiently large n, under some assumptions, with high probability,

- the optimal graph (from g^{opt}) $=$ the true graph
- the early-stop graph (from g^{early}) $=$ the true graph, when $\tau^{\mathrm{early}} = o(p \wedge \lambda_n)$

Statistical guarantees

Theorem. For sufficiently large n, under some assumptions, with high probability,

- the optimal graph (from g^{opt}) = the true graph
- the early-stop graph (from g^{early}) = the true graph, when $\tau^{\mathrm{early}} = o(p \wedge \lambda_n)$

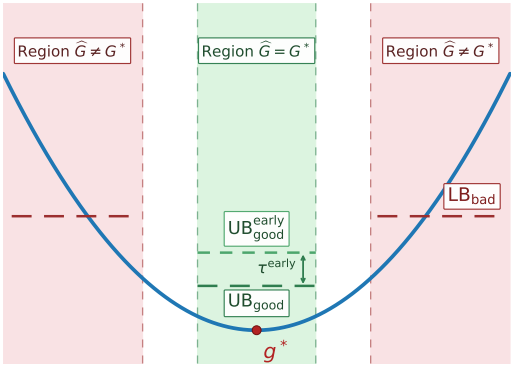
Assumptions.

- Model sparsity: $s^\star \leq s_n$
- Regularity conditions (e.g., beta-min, functional compatibility)
- Function classes for $\{f_{kj}\}$:
  - Sobolev function class for $\{f_{kj}\}$: $\|f_{kj} - \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}\|_\infty \lesssim R_n^{-\eta}$
  - Basis functions $\{b_{rkj}\}$ are orthogonal and bounded
  - $R_n \lesssim \sqrt{n/\log p}$
- A proper choice of λ_n: $\lambda_n \asymp (p + s_n R_n) \log p / n + s_n R_n^{-2\eta} + \gamma_n \|\beta^\star\|_2^2$

Statistical guarantees

Theorem. For sufficiently large n, under some assumptions, with high probability,

- the optimal graph (from g^{opt}) = the true graph
- the early-stop graph (from g^{early}) = the true graph, when $\tau^{\mathrm{early}} = o(p \wedge \lambda_n)$



- Region separation:
  Upper bound (Region $\widehat{G} = G^\star$)
  $<$ Lower bound (Region $\widehat{G} \neq G^\star$)
- assumptions $\rightarrow$ separation gap
- $\mathrm{obj}(g^{\mathrm{early}}) \leq \mathrm{obj}(g^{\mathrm{opt}}) + \tau^{\mathrm{early}}$

Experiments

MIP CP: our structure-aware MIP formulation

MIP Vanilla: the original MIP $\longrightarrow$ off-the-shelf solver

	MIP CP		MIP Vanilla		CAM		NPVAR	
Dataset.p.s*	Time	SHD	Time	SHD	Time	SHD	Time	SHD
dsep.6.6	0.10	0.3	0.19	0.3	1.62	2.3	0.14	8.5
asia.8.8	0.22	0.5	0.49	0.5	2.38	0.0	0.30	12.1
bowling.9.11	0.32	1.1	7.33	1.1	2.93	1.6	0.42	14
inssmall.16.25	2.18	5.1	301.1*	6.9	7.51	12.5	2.07	36.2
insurance.27.52	56.17	8.3	602.5*	12.0	15.58	16.7	15.3	82

* Time limit reached

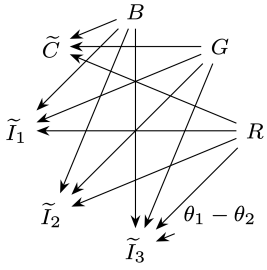
Takeaway: MIP CP is

- competitive speed-wise
- and much more accurate

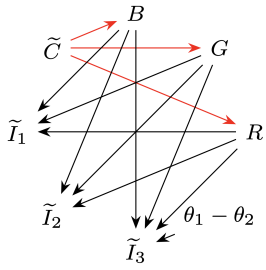
We can solve a DAG of 50 nodes (with $\tau^{\text{early}} > 0$) within 20 seconds and SHD=11

Real data: the causal chamber

- computer-controlled devices (light & wind tunnels)
- a testbed for causal discovery algorithms
- physical data generation mechanism:
  - manipulate and measure variables
  - the **ground truth** is known



(a) Ground truth



(b) MIP estimation

Method	SHD
NPVAR	14
NoTears	32
CAM	28
MIP	6

(c) Comparison

Red edges $\rightarrow$: incorrect discoveries

Takeaway: Our MIP

- captures the correct skeleton
- captures the correct non-linear edge ($\theta_1 - \theta_2 \rightarrow \tilde{I}_3$)
- is the most accurate

Summary and outlook

Summary: A convex MIP framework for non-linear DAG learning

- without equal-variance assumption
- background knowledge
- optimality and statistical guarantees

Outlook:

- more targeted estimation via constraints in MIP
- MIP for local discovery (e.g. parents for a target)
- uncertainty quantification/ confidence regions over DAGs

Thanks for listening!

Partially based on the article:
`https://arxiv.org/abs/2511.21126`