Block Successive Convex Approximation for Concomitant Linear DAG Estimation

Seyed Saman Saboksayr*, Gonzalo Mateos*, and Mariano Tepper†

*Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA

†Intel Labs, Hillsboro, OR, USA

{ssaboksa, gmateosb}@ur.rochester.edu, and mariano.tepper@intel.com

Overview of this work

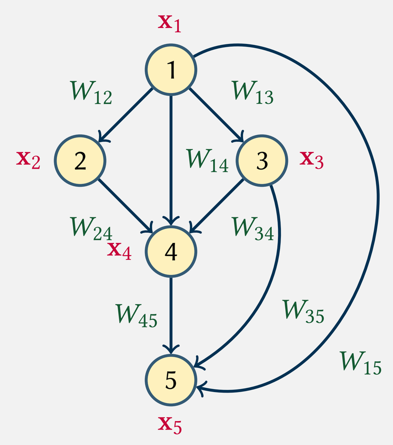
- **Goal:** Learning DAG structure from observational data
- Recently proposed Concomitant Linear DAG Estimation (CoLiDE) framework
  - ⇒ Jointly estimate DAG structure along with exogenous noise levels
  - ⇒ No parameter retuning needed and amenable to non-equal noise variances cases
- **Contribution:** Deriving efficient optimization algorithm, closed-form updates
  - ⇒ Leverages block successive convex approximation (BSCA) algorithm
  - ⇒ Providing a provably convergent sequence → Superior performance

What are DAGs and how to learn their connectivity structure?

- Directed graph $\mathcal{G}$ without cycles increasingly prominent in ML applications
  - ⇒ May encode causal relationships within complex systems
  - ⇒ Employ directed edges to link causes and their immediate effects
- Causal structure underlying a group of variables is often unknown
  - ⇒ Need to address the task of inferring DAGs from observational data
- Markovian linear structural equation model (SEM)

$$\mathbf{x}_i = \mathbf{w}_i^\top \mathbf{X} + \mathbf{z}_i, \quad \text{where} \quad \mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$$

  - ⇒ DAG adjacency matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_d] \in \mathbb{R}^{d \times d}$ collects the edge weights
  - ⇒ $\mathbf{z}_i \in \mathbb{R}^n$ is a vector of mutually independent, exogenous noises
  - ⇒ Ex: $\mathbf{x}_4 = \mathbf{w}_4^\top \mathbf{X} + \mathbf{z}_4 = W_{14}\mathbf{x}_1 + W_{24}\mathbf{x}_2 + W_{34}\mathbf{x}_3 + \mathbf{z}_4$



Problem statement: Given data $\mathbf{X}$ adhering to a linear SEM, learn the latent DAG $\mathcal{G} \in \mathbb{D}$, i.e., estimate its adjacency matrix $\mathbf{W}$ by minimizing the score function $\mathcal{S}$, namely

$$\min_{\mathbf{W}} \ \mathcal{S}(\mathbf{W}) \ \text{ subject to } \ \mathcal{G}(\mathbf{W}) \in \mathbb{D}$$

- Learning a DAG solely from observational data $\mathbf{X}$ is in general NP-hard
  - ⇒ Combinatorial acyclicity constraint $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$ is difficult to enforce
  - ⇒ Multiple DAGs can generate the same observational distribution

Concomitant linear DAG estimation (CoLiDE)

- Acyclicity characterization using nonconvex, smooth $\mathcal{H}(\mathbf{W}) : \mathbb{R}^{d \times d} \mapsto \mathbb{R}$
  - ⇒ Relax combinatorial constraint by enforcing $\mathcal{H}(\mathbf{W}) = 0 \iff \mathcal{G}(\mathbf{W}) \in \mathbb{D}$
  - ⇒ Ex: DAGMA formulation adopts $\mathcal{H}_{\text{ldet}}(\mathbf{W}, s) = d\log(s) - \log(\det(s\mathbf{I} - \mathbf{W} \circ \mathbf{W}))$
- Solve smooth, continuous optimization problem → $\min_{\mathbf{W}} \ \mathcal{S}(\mathbf{W})$ subject to $\mathcal{H}(\mathbf{W}) = 0$

CoLiDE-EV

- All exogenous variables $\mathbf{z}_1, \dots, \mathbf{z}_d$ in the linear SEM have equal variance (EV) σ^2
- Inspired by the smoothed concomitant lasso

$$\min_{\mathbf{W}, \sigma \geq \sigma_0} \underbrace{\frac{1}{2n\sigma}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \frac{d\sigma}{2} + \lambda\|\mathbf{W}\|_1}_{:=\mathcal{S}(\mathbf{W},\sigma)} \quad \text{subject to} \quad \mathcal{H}(\mathbf{W}) = 0$$

- Score $\mathcal{S}(\mathbf{W}, \sigma)$ is jointly convex, $(d\sigma)/2$ for consistency under Gaussianity
  - ⇒ λ decouples from σ as minimax optimality now requires $\lambda \asymp \sqrt{\log d / n}$
- Solve a sequence of unconstrained problems where $\mathcal{H}$ is viewed as a regularizer
- **Optimization:** Given a decreasing sequence $\mu_k \to 0$, at step k we solve

$$\min_{\mathbf{W}, \sigma \geq \sigma_0} \ \mu_k \left[\frac{1}{2n\sigma}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \frac{d\sigma}{2} + \lambda\|\mathbf{W}\|_1\right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)$$

  - ⇒ Limit $\mu_k \to 0$ is guaranteed to yield a DAG
  - ⇒ Jointly estimates the noise level σ and the adjacency matrix $\mathbf{W}$ for each μ_k
- Fixing σ to its latest value and minimizing score function inexactly w.r.t. $\mathbf{W}$
  - ⇒ One iteration of gradient descent via the ADAM optimizer
- Updating σ in closed form given the latest $\mathbf{W}$ via $\hat{\sigma} = \max\left(\frac{1}{\sqrt{nd}}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F, \sigma_0\right)$

CoLiDE-NV

- Noise variables $\mathbf{z}_1, \dots, \mathbf{z}_d$ have non-equal variances (NV) $\sigma_1^2, \dots, \sigma_d^2$
- Mimicking the previous optimization approach, we propose CoLiDE-NV

$$\min_{\mathbf{W}, \Sigma \geq \Sigma_0} \ \mu_k \left[\frac{1}{2n}\text{Tr}\left((\mathbf{X} - \mathbf{W}^\top\mathbf{X})^\top \Sigma^{-1} (\mathbf{X} - \mathbf{W}^\top\mathbf{X})\right) + \frac{1}{2}\text{Tr}(\Sigma) + \lambda\|\mathbf{W}\|_1\right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)$$

  - ⇒ $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_d)$ is a diagonal matrix of noise standard deviations
- Per iteration cost is $\mathcal{O}(d^3)$, on par with state-of-the-art DAG learning methods

Optimization revisited: Block Successive Convex Approximation (BSCA)

CoLiDE-EV

- Fixing σ to its most up-to-date value σ_t the resulting composite subproblem is

$$\min_{\mathbf{W}} \ \underbrace{\left[\frac{\mu_k}{2n\sigma_t}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)\right.}_{:=f(\mathbf{W})} + \underbrace{\left.\lambda\mu_k\|\mathbf{W}\|_1\right]}_{:=g(\mathbf{W})}$$

  - ⇒ $g(\mathbf{W})$ is convex but not smooth, while $f(\mathbf{W})$ is smooth but non-convex
- Quadratic approximation of $f(\mathbf{W})$ around the previous iterate $\mathbf{W}_{t-1}$

$$\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1}) := \langle \mathbf{W} - \mathbf{W}_{t-1}, \nabla f(\mathbf{W}_{t-1})\rangle + \frac{L}{2}\|\mathbf{W} - \mathbf{W}_{t-1}\|_F^2$$

  - ⇒ Strictly convex for any positive scalar L
- Instead of solving the original $\mathbf{W}$ subproblem, we can minimize the approximation

$$\bar{\mathbf{W}}_t = \underset{\mathbf{W}}{\text{argmin}} \ \left[\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1}) + \lambda\mu_k\|\mathbf{W}\|_1\right]$$

- Given the proximal operator of $g(\mathbf{W})$, closed-form update of $\bar{\mathbf{W}}_t$ is

$$\bar{\mathbf{W}}_t = \mathcal{T}_{\mu_k\lambda}\left(\mathbf{W}_{t-1} + \frac{\mu_k}{\sigma_t n}\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \mathbf{W}_{t-1}) - 2(s_k\mathbf{I} - \mathbf{W}_{t-1} \circ \mathbf{W}_{t-1})^{-\top} \circ \mathbf{W}_{t-1}\right)$$

  - ⇒ Soft-thresholding operator $\mathcal{T}_\alpha(x) = \max(|x| - \alpha, 0)\,\text{sign}(x)$
- **Challenge:** $\nabla\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1})$ is not Lipschitz continuous
  - ⇒ $\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1})$ is not guaranteed to be a global upper bound of $f(\mathbf{W})$
- We update the DAG adjacency matrix as

$$\mathbf{W}_t = \mathbf{W}_{t-1} + \gamma_t(\bar{\mathbf{W}}_t - \mathbf{W}_{t-1})$$

  - ⇒ Select $\gamma_t \in (0, 1]$ via the low-complexity Armijo rule

CoLiDE-NV

- Similar successive approximation methodology employed for the CoLiDE-NV cost

$$f(\mathbf{W}) := \frac{\mu_k}{2n}\text{Tr}\left((\mathbf{X} - \mathbf{W}^\top\mathbf{X})^\top\Sigma_t^{-1}(\mathbf{X} - \mathbf{W}^\top\mathbf{X})\right) + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)$$

- Again, the so-termed proximal linear approximation yields

$$\bar{\mathbf{W}}_t = \mathcal{T}_{\mu_k\lambda}\left(\mathbf{W}_{t-1} + \frac{\mu_k}{n}\mathbf{X}^\top\mathbf{X}\left[\mathbf{I} - \mathbf{W}\right]\Sigma_t^{-1} - 2(s_k\mathbf{I} - \mathbf{W}_{t-1} \circ \mathbf{W}_{t-1})^{-\top} \circ \mathbf{W}_{t-1}\right)$$

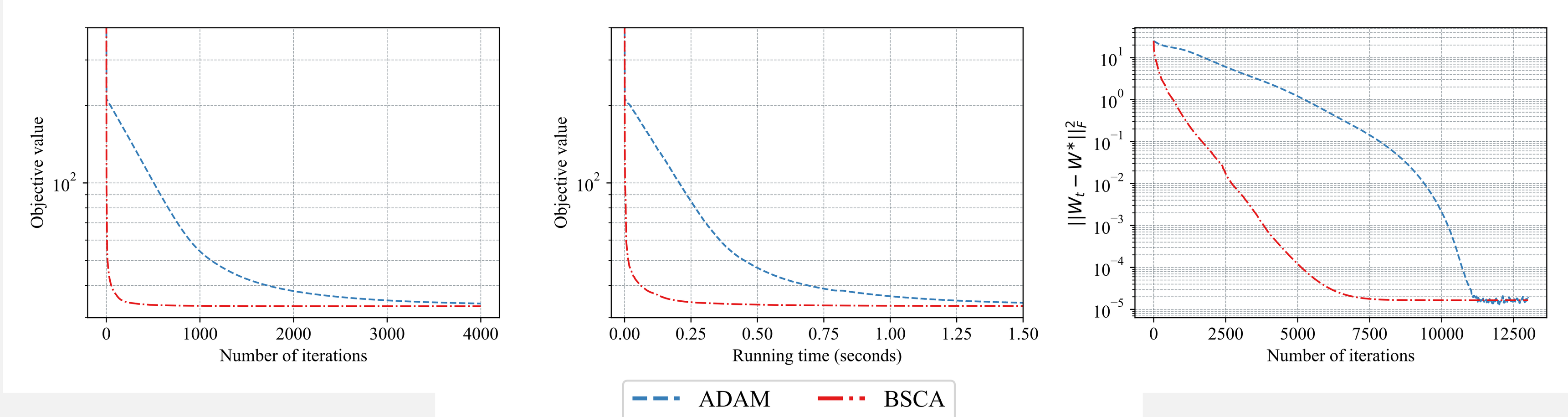
Convergence and complexity

- Every limit point of the BSCA sequence is a stationary point of original problem
- This comes with no order-wise penalty in computational complexity

Experimental evaluation

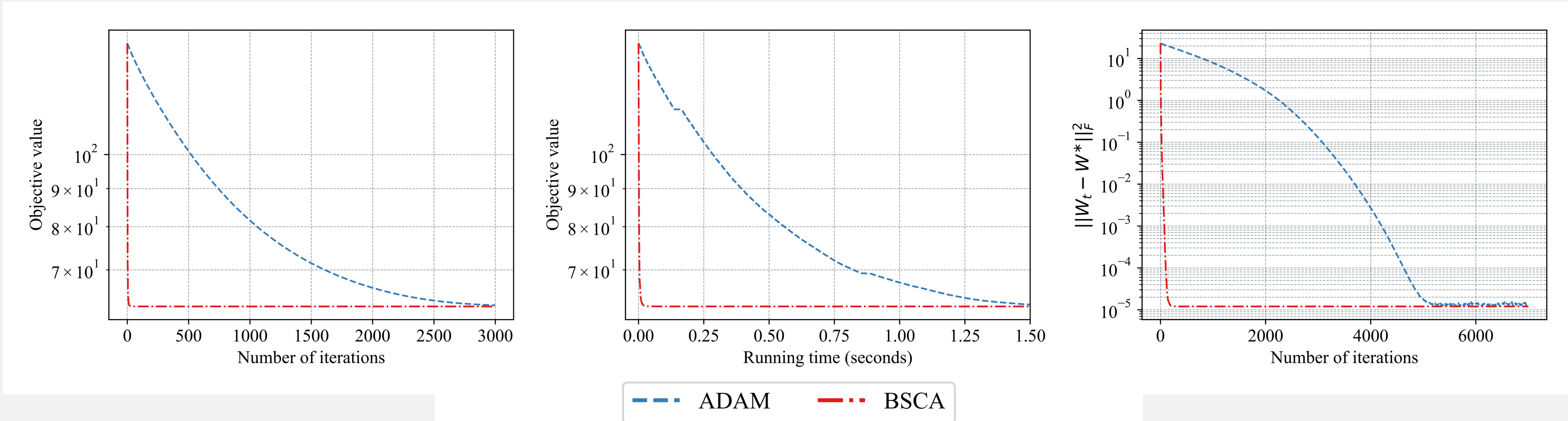
Equal variance experiments

- We consider a single step of the sequence where $\mu_k = 1$ and $s_k = 1$
- 50-node ER graph with 50 weighted edges $\mathcal{E} \in [-2, -0.5] \cup [0.5, 2]$
- Simulate $n = 1000$ samples considering Gaussian noise ($\sigma^2 = 1$) via linear SEM
- Optimal solution $\mathbf{W}^*$ is obtained by running the inexact BCD algorithm for 10^5 iterations



Non-equal variance experiments

- Noise variance of each node is uniformly drawn from $[0.5, 10]$
- 50-node ER graph with 50 weighted edges $\mathcal{E} \in [-1, -0.25] \cup [0.25, 1]$



References and GitHub page

S. S. Saboksayr, G. Mateos, and M. Tepper, "CoLiDE: Concomitant Linear DAG Estimation," In *Proc. Int. Conf. Learn. Representations*, 2024.

Y. Yang, M. Pesavento, Z.-Q. Luo, and B. Ottersten, "Inexact block coordinate descent algorithms for nonsmooth nonconvex optimization," *IEEE Trans. Signal Process.*, 2020.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with no tears: Continuous optimization for structure learning," In *Proc. Adv. Neural. Inf. Process. Syst.*, 2018.

K. Bello, B. Aragam, and P. Ravikumar, "DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization," In *Proc. Adv. Neural. Inf. Process. Syst.*, 2022.