Block Successive Convex Approximation for Concomitant Linear DAG Estimation

Seyed Saman Saboksayr
Dept. of Electrical and Computer Eng.
University of Rochester
ssaboksa@ur.rochester.edu

Gonzalo Mateos
Dept. of Electrical and Computer Eng.
University of Rochester
gmateosb@ur.rochester.edu

Mariano Tepper
Intel Labs
mariano.tepper@intel.com

***Abstract*—We develop a novel continuous optimization algorithm to recover latent directed acyclic graphs (DAGs) from observational (and possibly heteroscedastic) data adhering to a linear structural equation model (SEM). Our starting point is the recently proposed Concomitant Linear DAG Estimation (CoLiDE) framework, which advocates minimizing a sparsity-regularized convex score function augmented with a smooth, nonconvex acyclicity penalty. While prior work focused on score function design to jointly estimate DAG structure along with exogenous noise levels, optimization aspects were left unexplored. To bridge this gap, here we show that CoLiDE has a favorable structure amenable to optimization via a block successive convex approximation (BSCA) algorithm. We derive efficient, closed-form updates to refine the DAG adjacency matrix and noise variance estimates in a cyclic fashion. Although the acyclicity regularizer is devoid of a Lipschitz gradient and hence our approximation function is not a global upper bound of the original cost, a descent direction can be obtained via line search to yield a provably convergent sequence. Numerical tests showcase the superiority of the proposed BSCA iterations relative to the original (Adam-based) inexact block coordinate descent solver.**

***Index Terms*—Concomitant scale estimation, directed acyclic graph, successive convex approximation, topology inference.**

I. Introduction

Directed acyclic graphs (DAGs) are used to represent causal relationships among variables, where connections between causes and their immediate effects are encoded through directed edges. DAGs and associated Bayesian networks find applications in e.g., biology [1], [2], genetics [3], finance [4], and economics [5]. Since the causal structure underlying a collection of variables is typically unknown, inferring DAGs from nodal observations becomes a crucial task [6, Ch. 7]. However, learning a DAG solely from observational data (cf. interventional data) presents substantial computational challenges, primarily due to the well-documented difficulty of enforcing the combinatorial acyclicity constraint [7], [8]. Additionally, in general multiple DAGs can generate the same observational data distribution, making the identification of the true DAG nontrivial. This identifiability challenge may be pronounced when data are limited, or, when candidate graphs exhibit Markov equivalence; see e.g., [6].

DAG learning from observational data is an NP-complete problem [7], [8], and a recent flurry of successful approaches have advocated *continuous relaxations* leading to constrained optimization formulations [9]–[12]; see Section II for a problem statement. The selection of an appropriate score function plays a crucial role in effectively guiding continuous optimization techniques to recover the latent DAG. Regression-based score functions, such as sparsity-regularized ordinary least squares (LS) [9], [11], have been shown effective in high-dimensional settings characterized by data scarcity and model uncertainty; see also [13] for consistency results. However, lasso-type criteria to score DAGs in linear structural equation models (SEMs) typically rely on the assumption of homoscedasticity, i.e., exogenous noises have equal variances across variables. Additionally, they necessitate careful fine-tuning of the penalty parameter governing the trade-off between sparsity and data fidelity [14], [15]. To address these challenges, we recently proposed the Concomitant Linear DAG Estimation (CoLiDE) framework [12]. Leveraging ideas from concomitant scale estimation in sparse regression [16], [17], we put forth a novel convex score function for inference of DAGs in (possibly heteroscedastic) linear SEMs. CoLiDE exhibits significant improvements relative to existing state-of-the-art DAG learning methods; see [12] for further details.

While [12] focused on score function design to jointly estimate DAG structure along with exogenous noise levels (Section III), optimization aspects were left unexplored. Indeed, CoLiDE optimization therein relies on block coordinate descent (BCD) iterations, whereby the DAG subproblem is solved inexactly by running one iteration of the Adam optimizer [18]. While shown to be effective, this heuristic does not come with theoretical convergence guarantees. To bridge this gap, in Section IV we show that CoLiDE has a favorable structure amenable to optimization via a block successive convex approximation (BSCA) algorithm [19], [20]. We derive efficient, closed-form updates to refine the DAG adjacency matrix and noise variance estimates in a cyclic fashion. Although CoLiDE's log-determinant acyclicity regularizer [11] is devoid of a Lipschitz gradient and hence our approximation function is not a global upper bound of the original cost, a descent direction can be obtained via line search to yield a provably convergent sequence. All in all, our algorithmic contribution is envisioned to impact DAG learning and causal discovery.

II. Preliminaries and Problem Statement

Consider a directed graph (digraph) $\mathcal{G}\left(\mathcal{V}, \mathcal{E}, \mathbf{W}\right)$, where $\mathcal{V} = \{1, \ldots, d\}$ represents the set of vertices, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$

is the set of edges. The relationship between nodes is encoded in the adjacency matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_d] \in \mathbb{R}^{d \times d}$ where a non-zero edge weight W_{ij} implies a direct link from node i to node j. Let us assume the digraph $\mathcal{G}$ captures conditional independencies among the variables in the random vector $\mathbf{x} = [x_1, \dots, x_d]^\top \in \mathbb{R}^d$ and that it belongs to the space $\mathbb{D}$ of DAGs. If the joint distribution $\mathbb{P}(\mathbf{x})$ satisfies a Markov property over $\mathcal{G} \in \mathbb{D}$, it implies that each random variable x_i is solely dependent on its parents $\mathrm{PA}_i = \{j \in \mathcal{V} : W_{ji} \neq 0\}$ [6]. This work focuses on *linear* SEMs to generate such a probability distribution, where the relationship between each random variable and its parents is expressed as $x_i = \mathbf{w}_i^\top \mathbf{x} + z_i, \forall i \in \mathcal{V}$, where $\mathbf{z} = [z_1, \dots, z_d]^\top$ is a vector of mutually independent, exogenous noises, without any specific assumption on their distribution; see e.g., [6]. For a dataset $\mathbf{X} \in \mathbb{R}^{d \times n}$ consisting of n i.i.d. samples drawn from $\mathbb{P}(\mathbf{x})$, the linear SEM equations can be expressed in matrix form as $\mathbf{X} = \mathbf{W}^\top \mathbf{X} + \mathbf{Z}$.

Problem statement. Given the data matrix $\mathbf{X}$ generated by a linear SEM, the goal is to recover the underlying DAG $\mathcal{G} \in \mathbb{D}$. To estimate the adjacency matrix $\mathbf{W}$, one can solve

$$\min_{\mathbf{W}} \ \mathcal{S}(\mathbf{W}) \ \text{ subject to } \ \mathcal{G}(\mathbf{W}) \in \mathbb{D}, \tag{1}$$

where $\mathcal{S}(\mathbf{W})$ is a data-dependent score function to measure the quality of the candidate DAG. Regardless of the specific criterion employed, (1) is a hard non-convex problem due to the combinatorial acyclicity constraint $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$.

Typically, an effective score function incorporates a data fidelity term aligned with the SEM and regularization terms to encourage desired structural properties on the sought DAG. Since sparsity is a cardinal property of most real-world graphs, it is prudent to augment the widely-adopted ordinary LS loss with an ℓ_1-norm regularizer to yield $\mathcal{S}(\mathbf{W}) = \frac{1}{2n}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \lambda \|\mathbf{W}\|_1$, where $\lambda \geq 0$ is a tuning parameter that controls edge sparsity. This score function $\mathcal{S}(\mathbf{W})$ shares similarities with the *multi-task* variant of lasso regression [21], particularly when the response and design matrices coincide. The optimal rates for lasso are contingent on selecting $\lambda \asymp \sigma\sqrt{\log d/n}$ [22], [23], but the exogenous noise variance σ^2 is rarely known in practice. This challenge is compounded in heteroscedastic settings, where one must adopt a *weighted* LS score [13] to mitigate bias incurred in most DAG learning methods that use ordinary LS. Acknowledging these limitations, we proposed a novel LS-based score function to facilitate *joint* estimation of the DAG and noise levels [12]. Next, we briefly review the CoLiDE framework and then present our algorithmic innovations in Section IV.

III. Concomitant Linear DAG Estimation

A recent trend to handle the combinatorial constraint $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$, is to leverage non-convex, smooth functions $\mathcal{H} : \mathbb{R}^{d \times d} \mapsto \mathbb{R}$ of the adjacency matrix. These functions are chosen such that their zero level set is $\mathbb{D}$. Consequently, the DAG learning problem (1) can be relaxed by enforcing $\mathcal{H}(\mathbf{W}) = 0$ instead of $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$, paving the way for standard *continuous* optimization algorithms [9], [11], [24], [25].

In this context, CoLiDE introduces a novel convex score function for linear DAG estimation, incorporating concomitant estimation of scale parameters [12]; see [15], [17], [26], [27] for linear regression counterparts that inspired our work. CoLiDE exhibits robust DAG estimation performance in heteroscedastic settings and effectively decouples the sparsity parameter λ from the exogenous noise level σ.

CoLiDE-EV. Suppose all exogenous noise variables $z_1, \dots, z_d$ in the linear SEM have equal variance (EV) σ^2. To simultaneously estimate the DAG adjacency matrix $\mathbf{W}$ and the noise scale σ, our idea is to solve (see [12] for further details):

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

The weighted, regularized LS score function $\mathcal{S}(\mathbf{W}, \sigma)$ is jointly convex in $\mathbf{W}$ and σ, drawing inspiration from the robust linear regression work of [16]. Indeed, Huber pointed out that the inclusion of the linear term $d\sigma/2$ yields an estimator $\hat{\sigma}$ that is consistent under Gaussianity. Due to the rescaled residuals, the tuning parameter λ in (2) becomes independent of σ for minimax optimality, namely $\lambda \asymp \sqrt{\log d/n}$ [23]. With regards to the acyclicity function, we chose $\mathcal{H}_{\text{ldet}}(\mathbf{W}, s) = d\log(s) - \log(\det(s\mathbf{I} - \mathbf{W} \circ \mathbf{W}))$ [11], where $s \in \mathbb{R}$ and $\circ$ denotes Hadamard product. $\mathcal{H}_{\text{ldet}}$ has been shown to possess favorable gradient properties, along with several other desirable properties outlined in [11, Section 3.2].

To solve (2), we minimize a series of unconstrained problems wherein $\mathcal{H}_{\text{ldet}}$ is dualized and treated as a regularizer. For a sequence $\mu_k \to 0$, at step k COLIDE-EV solves

$$\min_{\mathbf{W}, \sigma \geq \sigma_0} \ \mu_k \left[\frac{1}{2n\sigma}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \frac{d\sigma}{2} + \lambda\|\mathbf{W}\|_1\right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k), \tag{3}$$

where $\mu_k, s_k > 0$ are hyperparameters and $\sigma_0 = \frac{\|\mathbf{X}\|_F}{\sqrt{dn}} \times 10^{-2}$. Decreasing μ_k amplifies the influence of the acyclicity function, and the limit $\mu_k \to 0$ is guaranteed to yield a DAG. This methodology resembles the central path of a barrier method, and in practice it proves to be more effective than alternatives such as the augmented Lagrangian method [11]. We typically select $\mu_k \in \{1, 0.1, 0.01, 0.001\}$ and $s_k \in \{1, 0.9, 0.8, 0.7\}$.

For each μ_k in the sequence, CoLiDE-EV employs (inexact) BCD iterations to jointly solve for the noise level σ and $\mathbf{W}$. This cyclic strategy entails keeping σ fixed at its most recent value and minimizing (3) inexactly w.r.t. $\mathbf{W}$, followed by a closed-form update of σ given the latest $\mathbf{W}$, namely

$$\hat{\sigma} = \max\left(\sqrt{\mathrm{Tr}\left((\mathbf{I} - \mathbf{W})^\top \mathrm{cov}(\mathbf{X})(\mathbf{I} - \mathbf{W})\right)/d}, \sigma_0\right), \tag{4}$$

where $\mathrm{cov}(\mathbf{X}) := \frac{1}{n}\mathbf{X}\mathbf{X}^\top$ is the sample covariance matrix. There are various alternatives to inexactly solve the $\mathbf{W}$ subproblem using first-order methods. Computational considerations motivated running a single Adam step to refine $\mathbf{W}$, which led to good empirical performance but without offering theoretical convergence guarantees [12].

CoLiDE-NV. Consider now the challenging problem of learning DAGs in heteroscedastic scenarios, where noise variances $\sigma_1^2, \ldots, \sigma_d^2$ are non-equal (NV). Building on the generalized concomitant multi-task lasso [14] and emulating the optimization approach for the EV setting discussed earlier, the CoLiDE-NV estimator is formulated as follows

$$\min_{\mathbf{W}, \mathbf{\Sigma} \geq \mathbf{\Sigma}_0} \; \mu_k \left[\frac{1}{2n} \operatorname{Tr}\left((\mathbf{X} - \mathbf{W}^\top \mathbf{X})^\top \mathbf{\Sigma}^{-1} (\mathbf{X} - \mathbf{W}^\top \mathbf{X}) \right) \right.$$
$$\left. + \frac{1}{2} \operatorname{Tr}(\mathbf{\Sigma}) + \lambda \|\mathbf{W}\|_1 \right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k). \quad (5)$$

Note that $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$ is a diagonal matrix of exogenous noise *standard deviations* (hence not a covariance matrix). Once more, we set $\mathbf{\Sigma}_0 = \sqrt{\operatorname{diag}\left(\operatorname{cov}(\mathbf{X})\right)} \times 10^{-2}$, where $\sqrt{(\cdot)}$ is meant to be taken element-wise. A closed form solution for $\mathbf{\Sigma}$ given $\mathbf{W}$ is also readily obtained,

$$\hat{\mathbf{\Sigma}} = \max\left(\sqrt{\operatorname{diag}\left((\mathbf{I} - \mathbf{W})^\top \operatorname{cov}(\mathbf{X}) (\mathbf{I} - \mathbf{W}) \right)}, \mathbf{\Sigma}_0 \right). \quad (6)$$

Both CoLiDE variants incur a per iteration cost of $\mathcal{O}(d^3)$, similar to other state-of-the-art DAG learning methods [9]–[11]. Accordingly, CoLiDE facilitates joint estimation of the adjacency matrix $\mathbf{W}$ and $\mathbf{\Sigma}$, in more general settings and with only marginal added complexity compared to the task of determining the DAG structure alone.

IV. Block Successive Convex Approximation

As discussed in the previous section, CoLiDE optimization entails solving a sequence of problems (3) [or (5) in the NV case] for few decreasing values of μ_k. Here we derive an efficient and provably convergent BSCA algorithm to solve (3) for each k, by bringing to bear the advances in [19].

Just like the original BCD algorithm in [12], we will update $\mathbf{W}$ and σ in a cyclic fashion. With $t = 0, 1, 2, \ldots$ denoting iterations, for fixed $\mathbf{W}_t$ the best-response update for σ is given by (4), just as before. The focus then shifts to the update of $\mathbf{W}$, when σ is fixed to its most up-to-date value σ_t. The resulting composite subproblem is

$$\min_{\mathbf{W}} \Bigg[\underbrace{\frac{\mu_k}{2n\sigma_t} \|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)}_{:= f(\mathbf{W})} + \underbrace{\lambda \mu_k \|\mathbf{W}\|_1}_{:= g(\mathbf{W})} \Bigg], \quad (7)$$

where $g(\mathbf{W})$ is convex but not smooth, while $f(\mathbf{W})$ is smooth but nonconvex because of $\mathcal{H}_{\text{ldet}}$. Since minimizing (7) is non trivial, the BSCA approach instead advocates optimizing a sequence of successively refined approximation subproblems that are much easier to solve. To this end, at iteration t we let

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

be the quadratic approximation of $f(\mathbf{W})$ around the previous iterate $\mathbf{W}_{t-1}$, which is strictly convex for any positive scalar L (we henceforth let $L = 1$ for notational simplicity) and satisfies the technical conditions (A1)-(A4) in [19]. Now, instead of solving the original subproblem (7), we can minimize the approximation

Algorithm 1: BSCA Algorithm for CoLiDE (step k)

```
Input prior $\mathbf{W}_{k-1}^*$, $\operatorname{cov}(\mathbf{X})$, parameters $\mu_k, s_k, \alpha, \beta, \lambda$
Initialize $\mathbf{W}_0 = \mathbf{W}_{k-1}^*$
for $t = 1, 2, \ldots,$ do
    Compute $\bar{\mathbf{W}}_t$ via (9) for EV or (12) for NV
    Compute stepsize $\gamma_t$ via the Armijo rule
    Update $\mathbf{W}_t = \mathbf{W}_{t-1} + \gamma_t(\bar{\mathbf{W}}_t - \mathbf{W}_{t-1})$
    Update $\sigma_t$ via (4) for EV or $\mathbf{\Sigma}_t$ via (6) for NV
end
Output DAG $\mathbf{W}_k^* := \mathbf{W}_t$ and noise scale $\sigma_t$ or $\mathbf{\Sigma}_t$
```

$$\bar{\mathbf{W}}_t = \operatorname*{argmin}_{\mathbf{W}} \left[\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1}) + \lambda \mu_k \|\mathbf{W}\|_1 \right]. \quad (8)$$

Unlike (7), the solution of (8) is given in closed form and it boils down to evaluating the proximal operator of $g(\mathbf{W})$, i.e.,

$$\bar{\mathbf{W}}_t = \mathcal{T}_{\mu_k \lambda}\big(\mathbf{W}_{t-1} + \frac{\mu_k}{\sigma_t} \operatorname{cov}(\mathbf{X})(\mathbf{I} - \mathbf{W}_{t-1})$$
$$- 2(s_k \mathbf{I} - \mathbf{W}_{t-1} \circ \mathbf{W}_{t-1})^{-\top} \circ \mathbf{W}_{t-1} \big), \quad (9)$$

where $\mathcal{T}_\alpha(x) = \max(|x| - \alpha, 0) \operatorname{sign}(x)$ is the soft-thresholding operator that we apply element-wise.

An unique aspect of our DAG learning problem is that $\nabla \tilde{f}(\mathbf{W}, \mathbf{W}_{t-1})$ is *not* Lipschitz continuous because of the log-determinant acyclicity function $\mathcal{H}_{\text{ldet}}$. Hence, the quadratic approximation function $\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1})$ is not guaranteed to be a global upper bound of $f(\mathbf{W})$. For this reason, as suggested in [19] our idea is to update the DAG adjacency matrix as

$$\mathbf{W}_t = \mathbf{W}_{t-1} + \gamma_t (\bar{\mathbf{W}}_t - \mathbf{W}_{t-1}), \quad (10)$$

where $\gamma_t \in (0, 1]$ is a stepsize chosen via line search. In our implementation, we select γ_t via the low-complexity Armijo rule with its usual scalar parameters $\alpha, \beta \in (0, 1)$ [28]; see [19] for a detailed implementation in the BSCA context.

Moving on to CoLiDE-NV for completeness, a similar successive approximation methodology can be employed to tackle (5). Supposing $\mathbf{\Sigma}$ is fixed to $\mathbf{\Sigma}_t$ obtained via the best-response update in (6), the $\mathbf{W}$ subproblem minimizes [cf. (5)]

$$f(\mathbf{W}) := \frac{\mu_k}{2n} \operatorname{Tr}\left((\mathbf{X} - \mathbf{W}^\top \mathbf{X})^\top \mathbf{\Sigma}_t^{-1} (\mathbf{X} - \mathbf{W}^\top \mathbf{X}) \right)$$
$$+ \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k). \quad (11)$$

Once more, we form a quadractic approximation $\tilde{f}(\mathbf{W}, \mathbf{W}_{t-1})$ around $\mathbf{W}_{t-1}$, and minimize (8) instead of (11) resulting in the so-termed proximal linear approximation

$$\bar{\mathbf{W}}_t = \mathcal{T}_{\mu_k \lambda}\big(\mathbf{W}_{t-1} + \mu_k \operatorname{cov}(\mathbf{X}) \left[\mathbf{I} - \mathbf{W} \right] \mathbf{\Sigma}_t^{-1}$$
$$- 2(s_k \mathbf{I} - \mathbf{W}_{t-1} \circ \mathbf{W}_{t-1})^{-\top} \circ \mathbf{W}_{t-1} \big). \quad (12)$$

Just like for CoLiDE-EV, the DAG adjacency matrix is finally updated via (10). The BSCA algorithm iterations for both versions of CoLiDE are tabulated under Algorithm 1.

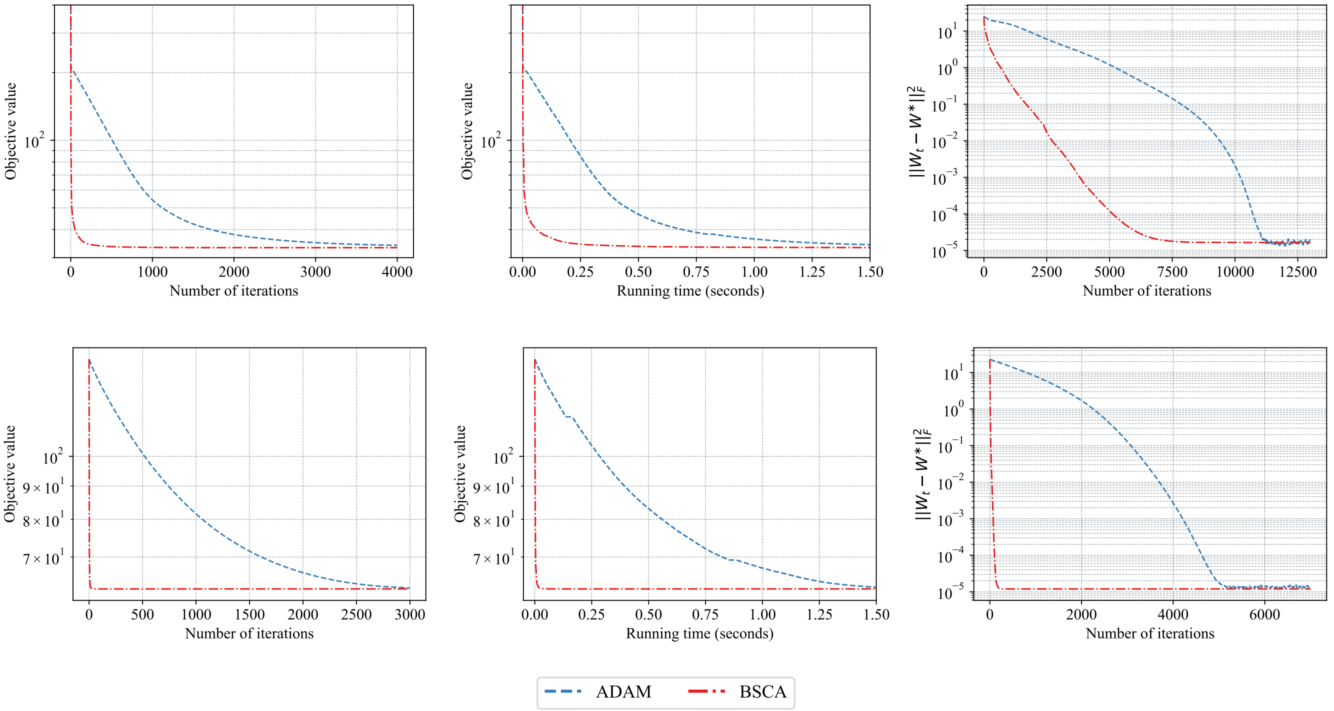

Fig. 1. Assessing the empirical convergence properties of the proposed BSCA algorithm in comparison to Adam-based inexact BCD in [12]. The top row considers scenarios where the noise variance is equal across all nodes (EV), while the bottom row assumes different noise variances across nodes (NV). BSCA exhibits superior performance across all experiments and metrics (number of iterations or wall-clock time to convergence).

Unlike the Adam-based inexact BCD heuristic in [12], by virtue of [19, Th. 1] it follows that every limit point of the BSCA sequence generated by Algorithm 1 is a stationary point of (3) [or (5) in the NV case]. This comes with no order-wise penalty in computational complexity.

V. Experimental Results

We conduct numerical experiments to analyze the performance of the proposed BSCA algorithm. CoLiDE's merits in terms of recovering high-quality DAGs have been well documented [12]. For this reason, here we will exclusively focus on algorithmic performance, with no examination of solution quality. As a baseline we consider the Adam-based inexact BCD heuristic in [12], and compare its empirical convergence properties against Algorithm 1. To this end, we generate a $d = 50$-node random Erdős-Rényi (ER) graph with 50 edges. Subsequently, we generate $n = 1000$ i.i.d. samples using a linear SEM, assuming the exogenous noises are Gaussian distributed. Given our focus on assessing the algorithm's performance in solving the optimization problem, we consider a single step of the sequence where $\mu_k = 1$ and $s_k = 1$. Following the CoLiDE guidelines in [12], we set $\lambda = 0.05$. As for α and β, we opt for typical values without engaging in hyperparameter tuning, choosing 0.01 and 0.25, respectively. To assess the empirical convergence performance, the columns of Figure 1 depict the objective value versus iterations, the objective value versus the running time, and the difference between the estimated DAG $\mathbf{W}_t$ at iteration t and the optimal solution $\mathbf{W}^*$. The latter is obtained by running the inexact BCD algorithm for 10^5 iterations. Note that $\mathbf{W}^*$ is the optimal solution of (3) [or (5)] for the given μ_k, and it may differ from the ground truth DAG.

For the EV case, we assume the noise variances are all $\sigma^2 = 1$, and the DAG edge weights are drawn uniformly at random from $[-2, -0.5] \cup [0.5, 2]$. As illustrated in the top row of Figure 1, the proposed BSCA algorithm consistently outperforms the Adam-based baseline in terms of convergence rate and wall-clock time. We follow the same procedure for CoLiDE-NV, where the noise variance of each node is randomly drawn from $[0.5, 10]$, and edge weights are chosen from $[-1, -0.25] \cup [0.25, 1]$. The bottom row of Figure 1 depicts the results for the heteroscedastic case, once again showcasing the superiority of Algorithm 1.

VI. Conclusion

We propose algorithmic advances for the CoLiDE framework to learn DAG topologies from linear SEM observations, while simultaneously estimating the exogenous noise levels for added robustness. Our contribution is to develop a novel BSCA algorithm with efficient updates that are given in closed form. Relative to the current Adam-based inexact BCD heuristic for this problem, the novel iterations are provably convergent and incur marginal added complexity stemming from the line search required to determine a suitable stepsize. The BSCA algorithm's superior convergence properties are demonstrated through preliminary experiments for both EV and NV settings.

References

[1] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[2] P. J. Lucas, L. C. Van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health-care," *Artif. Intell. Med.*, vol. 30, no. 3, pp. 201–214, 2004.

[3] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin *et al.*, "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[4] A. D. Sanford and I. A. Moosa, "A Bayesian network structure for operational risk modelling in structured finance operations," *J. Oper. Res. Soc.*, vol. 63, pp. 431–444, 2012.

[5] O. Pourret, P. Na, B. Marcot *et al.*, *Bayesian Networks: A Oractical Guide to Applications*. John Wiley & Sons, 2008.

[6] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[7] D. M. Chickering, "Learning Bayesian networks is NP-complete," *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130, 1996.

[8] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *J. Mach. Learn. Res.*, vol. 5, pp. 1287–1330, 2004.

[9] 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.*, vol. 31, 2018.

[10] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and DAG constraints for learning linear DAGs," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 33, 2020, pp. 17 943–17 954.

[11] 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.*, vol. 35, 2022, pp. 8226–8239.

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

[13] P.-L. Loh and P. Bühlmann, "High-dimensional learning of linear causal networks via inverse covariance estimation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3065–3105, 2014.

[14] M. Massias, O. Fercoq, A. Gramfort, and J. Salmon, "Generalized concomitant multi-task lasso for sparse multimodal regression," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 998–1007.

[15] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon, "Efficient smoothed concomitant Lasso estimation for high dimensional regression," in *J. Phys.: Conf. Ser.*, vol. 904, 2017, p. 012006.

[16] P. J. Huber, *Robust Statistics*. New York: John Wiley & Sons Inc., 1981.

[17] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemp. Math.*, vol. 443, no. 7, pp. 59–72, 2007.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[19] Y. Yang, M. Pesavento, Z.-Q. Luo, and B. Ottersten, "Inexact block coordinate descent algorithms for nonsmooth nonconvex optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 947–961, 2020.

[20] M. Razaviyayn, M. Hong, Z.-Q. Luo, and J.-S. Pang, "Parallel successive convex approximation for nonsmooth nonconvex optimization," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2014, p. 1440–1448.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc., B: Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[22] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, vol. 37, pp. 1705–1732, 2009.

[23] X. Li, H. Jiang, J. Haupt, R. Arora, H. Liu, M. Hong, and T. Zhao, "On fast convergence of proximal algorithms for SQRT-Lasso optimization: Don't worry about its nonsmooth loss function," in *Proc. Conf. Uncert. Artif. Intell.*, 2020, pp. 49–59.

[24] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.

[25] D. Wei, T. Gao, and Y. Yu, "DAGs with no fears: A closer look at continuous optimization for learning Bayesian networks," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 33, 2020, pp. 3895–3906.

[26] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[27] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[28] Y. Yang and M. Pesavento, "A unified successive pseudoconvex approximation framework," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3313–3328, 2017.