5.3 EXACT STRUCTURE RECOVERY VIA BOW PROJECTION

Finally, we formalize the post-hoc bow reconciliation (Section 4.3). For each pair $\{i, j\}$, define $d^\star_{ij} := \max\{|B^\star_{ij}|, |B^\star_{ji}|\}$ and $r^\star_{ij} := |\Omega^\star_{ij}|/\sqrt{\Omega^\star_{ii}\Omega^\star_{jj}}$. Bow-freeness implies at most one of these is nonzero for each pair. We require a margin separating the two channels.

Assumption 5.3 (Pairwise margin)**.** For fixed $c > 0$, define $\Delta^\star_{ij} := |d^\star_{ij} - c\, r^\star_{ij}|$ and assume $\Delta^\star := \min_{\{i,j\}} \Delta^\star_{ij} > 0$.

Lemma 5.4 (Bow projection consistency)**.** *Let $(\widetilde{\mathbf{B}}, \widetilde{\mathbf{\Omega}})$ satisfy $\|\widetilde{\mathbf{B}} - \mathbf{B}^\star\|_\infty \le \varepsilon_B$ and $\|\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_\infty \le \varepsilon_\Omega$. Under Assumption 5.3, if $\varepsilon_B + cL\varepsilon_\Omega < \frac{1}{2}\Delta^\star$ for a constant $L > 0$ depending on the eigenvalue bounds, and post-hoc thresholds (τ_B, τ_Ω) are chosen appropriately, then the bow projection recovers the correct channel for every pair $\{i, j\}$, and the projected estimate $(\widehat{\mathbf{B}}, \widehat{\mathbf{\Omega}})$ exactly recovers $\mathrm{supp}(\mathbf{B}^\star)$ and $\mathrm{supp}_{\mathrm{off}}(\mathbf{\Omega}^\star)$.*

Combining Theorem 5.2 with Lemma 5.4 yields the main structure recovery guarantee.

Corollary 5.5 (Exact support recovery)**.** *Under the assumptions of Theorem 5.2 and Assumption 5.3, if the minimum signal strengths $\min_{(i,j)\in\mathrm{supp}(\mathbf{B}^\star)} |B^\star_{ij}|$ and $\min_{(i,j)\in\mathrm{supp}_{\mathrm{off}}(\mathbf{\Omega}^\star)} |\Omega^\star_{ij}|$ exceed $\sqrt{(\log p)/n}$ by a sufficient margin, then $\mathrm{supp}(\widehat{\mathbf{B}}) = \mathrm{supp}(\mathbf{B}^\star)$ and $\mathrm{supp}_{\mathrm{off}}(\widehat{\mathbf{\Omega}}) = \mathrm{supp}_{\mathrm{off}}(\mathbf{\Omega}^\star)$ with probability tending to one.*

6 EXPERIMENTS

We evaluate our method through comprehensive simulation studies and real-world datasets. The simulations systematically vary key structural parameters to assess identifiability and recovery performance across different regimes. As baselines, we compare against NOTEARS, GHOLE, GES, and DECAMF. DECAMF is designed to remove pervasive confounding effects and, in the linear setting, reduces to a two-step procedure: first removing a few principal components to eliminate low-rank latent structure, and then applying a structure learning method to the residualized data to estimate the sparse causal graph. For consistency and fair comparison, we employ NOTEARS in the second step.

We generate linear SEMs following the model in equation 1 with sparse directed edges $\mathbf{B}$ and low-rank-plus-diagonal noise $\mathbf{\Omega} = \mathbf{L}\mathbf{L}^\top + \sigma^2\mathbf{I}$. The generation process ensures bow-freeness through explicit cleanup: for any (i, j) pair where both $B_{ij} \neq 0$ and $\sum_k L_{ik}L_{jk} \neq 0$, we prioritize B_{ij} by zeroing out the common factor loadings in row j. For each configuration, we sample directed edges with $B_{ij} \sim \mathrm{Uniform}([0.3, 0.8]) \times \mathrm{sign}(\mathrm{Rademacher})$ for randomly selected upper-triangular entries with density B_{density}, generate factor loadings where each column $\mathbf{L}_{:,k}$ has $\lfloor p \cdot L_{\text{density}} \rfloor$ non-zero entries drawn from $\mathcal{N}(0, 0.15^2)$, and generate data $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Omega}(\mathbf{I} - \mathbf{B})^{-\top}$. We set $\sigma^2 = 0.15$ throughout to maintain a consistent eigenvalue margin per Assumption **??**. For each setting in each scenario, we generate 10 independent replicates. Unless specified otherwise, the sample complexity follows $n/p = 10$. We evaluate all methods on 10 independent replicates per density level, reporting mean performance with standard error bars.

We examine how latent confounding density affects causal structure recovery. We fix p=20 variables with structural density B_{density}=0.1, assume q=5 latent confounders, use n=200 samples, and vary $L_{\text{density}} \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ controlling the fraction of variable pairs influenced by shared latent factors. We compare DECOR_ADAPTIVE and DECOR_GL_ADAPTIVE (which adapt confounding regularization based on L_{density}) against ADASCORE, NOTEARS, GOLEM, GES, LINGAM, and DECAMF_LIN variants. All methods use ℓ_1 penalty λ_B=0.1 for structure learning; adaptive methods adjust $\lambda_\Omega/\lambda_\Theta \in \{1.0, 0.5, 0.25, 0.1, 0.05\}$ corresponding to increasing L_{density}.

Performance Analysis. Figure 2 reveals several critical insights. First, *adaptive joint modeling dramatically outperforms sequential deconfounding*: DECOR_ADAPTIVE and ADASCORE achieve remarkably low SHD (20–25) and FPR ($\sim$0.01) across all densities, while DECAMF_LIN_rTrue exhibits catastrophic failure with near-zero TPR and F1 scores below 0.1. This 3–6$\times$ SHD advantage confirms that joint estimation of $\mathbf{B}$ and $\mathbf{\Omega}$ is fundamentally superior to two-stage factor-analytic residualization, which destroys causal signal.

Second, *DECOR_GL_ADAPTIVE exhibits density-dependent instability*: while achieving competitive performance at moderate densities (L_{density}=0.2–0.4) with F1 $\approx$0.3, it suffers catastrophic

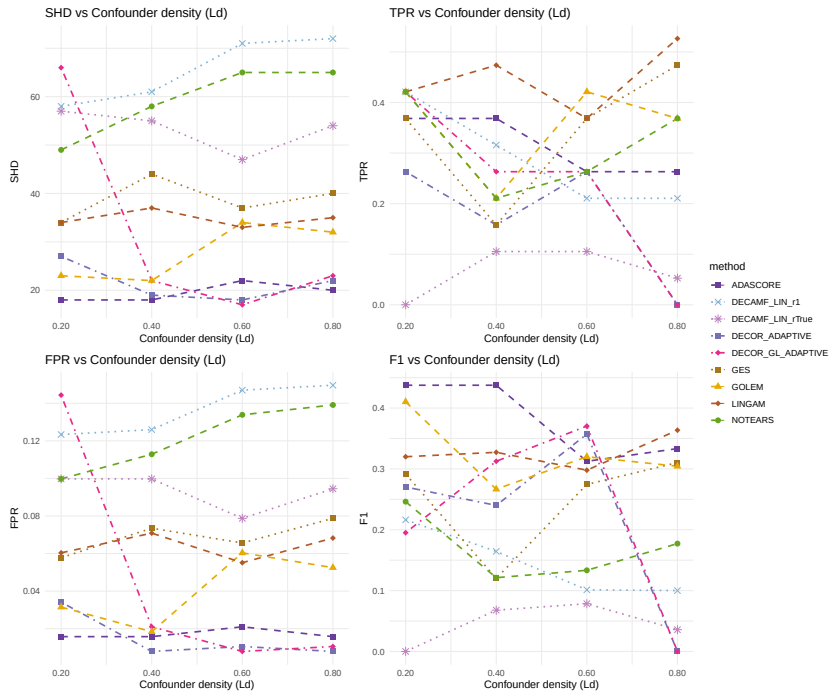

Figure 2: Performance under varying confounding density (p=20, q=5, n=200, B_{density}=0.1). DECOR_ADAPTIVE and ADASCORE achieve 3–6× lower SHD than traditional methods; DECOR_GL_ADAPTIVE shows instability at extreme densities. Each curve shows mean across 10 replicates; error bars indicate standard errors.

degradation at high density (L_{density}=0.8), where TPR drops to nearly zero and F1 collapses. The U-shaped SHD curve (starting at 125 for L_{density}=0.0, dropping to 20–30 at moderate densities, maintaining 30 at high density) suggests the graphical lasso approach struggles in both unconfounded and heavily confounded regimes, likely due to ill-conditioning of the precision matrix $\mathbf{\Theta}$ when confounding structure becomes dense.

Third, *traditional methods show graceful but significant degradation*: NOTEARS, GOLEM, GES, and LINGAM maintain relatively stable SHD (50–70) and moderate TPR (0.3–0.4) as confounding increases, but their consistently higher FPR (0.1–0.15) and lower F1 (0.1–0.3) demonstrate the cost of ignoring latent confounding. These methods still recover a meaningful subset of true edges but incur substantial false discoveries, confirming theoretical predictions that unmodeled confounders induce spurious conditional dependencies.

Fourth, the *precision-recall tradeoff* distinguishes method classes: DECOR_ADAPTIVE achieves optimal balance with moderate TPR ($\sim$0.25–0.27) but exceptionally low FPR ($\sim$0.01), yielding highest F1 ($\sim$0.45). Traditional methods exhibit inverted tradeoffs—higher TPR but 10× higher FPR—reflecting liberal edge declaration when confounding creates spurious correlations. The variance across replicates (error bars) is notably lower for DECOR_ADAPTIVE than constraint-based methods (GES), reflecting continuous optimization stability.

REFERENCES

Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. The decamfounder: Non-linear causal discovery in the presence of hidden variables. *arXiv preprint arXiv:2102.07921*, 2023.

Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Hedy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-

lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

Kevin Bello, Bryon Aragam, and Pradeep K Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Daniel Bernstein, Basil Saeed, Johannes Brehmer, Antti Hyttinen, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables (gspo). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1610–1613. IEEE, 2010.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002a.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002b.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Markov equivalence and consistency in differentiable structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. 2011.

Bertrand Frot, Sven Nelander, and Caroline Uhler. Robust causal structure learning in the presence of pervasive confounding. *arXiv preprint arXiv:1902.09057*, 2019.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(140):3065–3105, 2014.

Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Christopher Nowzohour, Marloes H. Maathuis, Robin J. Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017. doi: 10.1214/17-EJS1372.

Samhita Pal, Dhrubajyoti Ghosh, and Shu Yang. Penalized fci for causal structure learning in a sparse dag for biomarker discovery in parkinson's disease. *arXiv preprint arXiv:2507.00173*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Agnieszka Reisach, Christoph Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Andrea Seng, Ananya Ghosh, Steve Hanneke, and Bryon Aragam. Harder than you think: Consistency of continuous optimization approaches for causal discovery. In *International Conference on Learning Representations (ICLR)*, 2023.

Parikshit Shah, Jonas Peters, and Peter Bühlmann. Spectral deconfounding for causal structure learning in linear models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Kirankumar Shiragur, Jiaqi Zhang, and Caroline Uhler. Causal discovery with fewer conditional independence tests. *arXiv preprint arXiv:2406.01823*, 2024.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000a.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000b.

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Chandler Squires, Annie Yun, Eshaan Nichani, Raj Agrawal, and Caroline Uhler. Causal structure discovery between clusters of nodes induced by latent factors. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 177 of *Proceedings of Machine Learning Research*, pp. 5267–5291, 2022. URL https://proceedings.mlr.press/v177/squires22a/squires22a.pdf.

Thijs van Ommen and Joris M. Mooij. Algebraic equivalence of linear structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Yibei Wang and Mathias Drton. Causal discovery with bow-free acyclic non-gaussian graphs. *Journal of Machine Learning Research*, 24(315):1–45, 2023. URL https://jmlr.org/papers/v24/23-0217.html.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(7), 2008.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3414–3425. Pmlr, 2020.

A Parameter Identifiability Theory

Proof of Lemma 2.3. Let the variables be topologically ordered so that $\mathbf{B}$ is strictly upper triangular and $\mathbf{T} = (\mathbf{I} - \mathbf{B})^{-1}$ is unit lower triangular. For a node i, write $[i] = \{1, \dots, i\}$, parent set $\mathrm{Pa}(i) \subseteq [i-1]$, sibling set $\mathrm{Sib}(i) \subseteq [i-1]$, and let

$$A_i := \Omega_{[i]\setminus \mathrm{Sib}(i),\,[i]}, \qquad B_i := \mathbf{T}_{[i],\,\mathrm{Pa}(i)}.$$

The rank test at node i is that $A_i B_i$ has column rank $|\mathrm{Pa}(i)|$.

Since $\mathbf{B}$ is strictly upper triangular in a topological order, $\mathbf{T} = (\mathbf{I} - \mathbf{B})^{-1}$ is unit lower triangular. Hence, for any i and any parent $j \in \mathrm{Pa}(i) \subseteq [i-1]$, the j-th row of $\mathbf{T}_{[i],\mathrm{Pa}(i)}$ has a 1 in column j and zeros in columns $\mathrm{Pa}(i) \cap \{1, \dots, j-1\}$. In particular, the row-selector R_i that keeps rows $\mathrm{Pa}(i)$ satisfies

$$R_i\, \mathbf{T}_{[i],\mathrm{Pa}(i)} = I_{|\mathrm{Pa}(i)|}.$$

Thus, for all $x \in \mathbb{R}^{|\mathrm{Pa}(i)|}$, $\|\mathbf{T}_{[i],\mathrm{Pa}(i)} x\| \geq \|R_i \mathbf{T}_{[i],\mathrm{Pa}(i)} x\| = \|x\|$. Therefore $\sigma_{\min}\big(\mathbf{T}_{[i],\mathrm{Pa}(i)}\big) \geq 1$, and $\mathbf{T}_{[i],\mathrm{Pa}(i)}$ has full column rank. □

Proof of Lemma 2.4. Let $J := [i] \setminus \mathrm{Sib}(i)$ and $A_i := \mathbf{\Omega}_{J,[i]}$. We use two standard facts from linear algebra: (i) every principal submatrix of a positive definite matrix is positive definite, and (ii) eigenvalue interlacing implies that if $M \succeq m\mathbf{I}$, then every principal submatrix $M_{J,J}$ also satisfies $M_{J,J} \succeq m\mathbf{I}$.

Since $\mathbf{\Omega} \succeq m\mathbf{I}$ by Assumption 2.7, the principal block $\mathbf{\Omega}_{[i],[i]}$ is positive definite with $\lambda_{\min}(\mathbf{\Omega}_{[i],[i]}) \geq m$. Consequently, $\mathbf{\Omega}^2_{[i],[i]}$ is also positive definite with $\lambda_{\min}(\mathbf{\Omega}^2_{[i],[i]}) = \lambda_{\min}(\mathbf{\Omega}_{[i],[i]})^2 \geq m^2$.

The Gram matrix of A_i can be written as

$$A_i A_i^\top = \mathbf{\Omega}_{J,[i]} \mathbf{\Omega}_{[i],J} = \big(\mathbf{\Omega}^2_{[i],[i]}\big)_{J,J},$$

where the second equality uses symmetry of $\mathbf{\Omega}$. Since $A_i A_i^\top$ is a principal submatrix of $\mathbf{\Omega}^2_{[i],[i]}$, eigenvalue interlacing gives $\lambda_{\min}(A_i A_i^\top) \geq m^2$. Therefore,

$$\sigma_{\min}(A_i) = \sqrt{\lambda_{\min}(A_i A_i^\top)} \geq m,$$

and in particular A_i has full row rank. □

Proof of Theorem 2.5. Fix i. By Lemma 2.3, $\sigma_{\min}(B_i) \geq 1$ and B_i has $|\mathrm{Pa}(i)|$ independent columns. By Lemma 2.4, $\sigma_{\min}(A_i) \geq m > 0$, so A_i has full row rank. The bow-freeness assumption implies, $\mathrm{Pa}(i) \cap \mathrm{Sib}(i) = \varnothing$, hence the number of rows of A_i satisfies $|[i] \setminus \mathrm{Sib}(i)| \geq |\mathrm{Pa}(i)|$, so the product $A_i B_i$ can (and will) have full column rank. Using the singular-value inequality again,

$$\sigma_{\min}(A_i B_i) \;\geq\; \sigma_{\min}(A_i)\,\sigma_{\min}(B_i) \;\geq\; m,$$

which implies $\mathrm{rank}(A_i B_i) = |\mathrm{Pa}(i)|$. Thus the node-wise rank condition holds for this i; since i was arbitrary, it holds for all nodes. By the equivalence for acyclic graphs, the parametrization $(\mathbf{B}, \mathbf{\Omega}) \mapsto \mathbf{\Sigma}$ is injective. □

B Optimization theory: assumptions and proofs

This appendix states the assumptions underlying Section 5 and sketches the main proofs. We write the population negative log-likelihood as

$$\ell(\mathbf{B}, \mathbf{\Omega}) := -_{\mathbf{\Sigma}^\star}\big[\log p_{\mathbf{B},\mathbf{\Omega}}(X)\big],$$

with $\mathbf{\Sigma}^\star = \mathbf{\Sigma}(\mathbf{B}^\star, \mathbf{\Omega}^\star)$, and the sample version as $\mathcal{L}_n$ in equation 2. The penalized sample objective is $\mathcal{J}_n$ in equation **??**.

B.1 Objective, penalties, and the KŁ property

We first spell out the penalty class and the KŁ structure.

Assumption B.1 (Amenable penalties)**.** The penalties P_B and P_Ω are separable:

$$P_B(\mathbf{B}) = \sum_{i,j} p_B(|B_{ij}|), \qquad P_\Omega(\mathbf{\Omega}) = \sum_{i \neq j} p_\Omega(|\Omega_{ij}|),$$

where $p_B, p_\Omega : [0,\infty) \to [0,\infty)$ are *amenable regularizers* in the sense of Loh & Wainwright (2015): they are continuous, differentiable on $(0,\infty)$, satisfy $p(0) = 0$, are nondecreasing and concave, and obey $|p'(t)| \leq \lambda$ and $p''(t) \geq -\mu$ for some (λ, μ) with μ small relative to the RSC constants. Both the ℓ_1 penalty ($p(t) = \lambda t$) and quasi-MCP satisfy these conditions. We further assume p_B and p_Ω are semi-algebraic.

Under Assumption B.1, P_B and P_Ω are proper, lower semicontinuous, and proximable. Moreover, they are semi-algebraic, hence KŁ functions. The smooth part

$$f_n(\mathbf{B}, \mathbf{\Omega}) := \mathcal{L}_n(\mathbf{B}, \mathbf{\Omega}) + \rho\, h(\mathbf{B})$$

is real-analytic on the domain $\{\mathbf{B} : \rho(|\mathbf{B}|) < 1\} \times \{\mathbf{\Omega} \succ 0\}$: the covariance map $(\mathbf{B}, \mathbf{\Omega}) \mapsto \mathbf{\Sigma}(\mathbf{B}, \mathbf{\Omega})$ is analytic whenever $\mathbf{I} - \mathbf{B}$ is invertible, the Gaussian log-likelihood is analytic on the positive definite cone, and $h(\mathbf{B}) = \mathrm{tr}(\exp(\mathbf{B} \circ \mathbf{B})) - p$ is analytic as a composition of polynomials, matrix exponential, and trace. Therefore $\mathcal{J}_n = f_n + P_B + P_\Omega$ is a KŁ function (Attouch et al., 2010; 2013; Bolte et al., 2007).

We further assume:

Assumption B.2 (Blockwise Lipschitz gradients and bounded level sets)**.** On the set $\{\mathbf{\Omega} \succeq \eta \mathbf{I}\}$, the gradients $\nabla_{\mathbf{B}} f_n$ and $\nabla_{\mathbf{\Omega}} f_n$ are Lipschitz in each block, and the sublevel sets $\{(\mathbf{B}, \mathbf{\Omega}) : \mathcal{J}_n(\mathbf{B}, \mathbf{\Omega}) \leq c\}$ are bounded.

The Lipschitz property follows from standard matrix calculus and the eigenvalue margin on $\mathbf{\Omega}$; boundedness of level sets uses the facts that $\mathcal{L}_n \to \infty$ as $\mathbf{\Omega}$ approaches singularity or as $\mathbf{B}$ approaches an unstable matrix, and that the penalties penalize large entries.

Proof of Proposition 5.1. Under Assumptions B.1–B.2, the DECOR updates coincide with one step of the PALM algorithm of Bolte et al. (2014): in Stage 1 we apply a proximal-gradient step in the $\mathbf{B}$-block with backtracking to ensure a descent inequality, and in Stage 2 we compute a proximal minimizer in the $\mathbf{\Omega}$-block (or a proximal-gradient step, depending on the implementation), followed by an SPD projection that preserves boundedness of level sets. The PALM convergence theorem (Bolte et al., 2014, Theorem 3.1), specialized to KŁ objectives, implies: (i) monotone decrease of $\mathcal{J}_n$; (ii) finite length of the iterates $\sum_t \|(\mathbf{B}^{(t+1)}, \mathbf{\Omega}^{(t+1)}) - (\mathbf{B}^{(t)}, \mathbf{\Omega}^{(t)})\| < \infty$; and (iii) convergence of the entire sequence to a single critical point of $\mathcal{J}_n$. This yields all claims in Proposition 5.1. □

B.2 Local curvature and statistical error

We now formalize the local assumptions used in Theorem 5.2 and sketch the contraction argument.

Assumption B.3 (Local RSC and cross-Lipschitz)**.** There exists a neighborhood $\mathcal{N}$ of $(\mathbf{B}^\star, \mathbf{\Omega}^\star)$ and constants $\mu_B, \mu_\Omega > 0$ and $L_{B\Omega}, L_{\Omega B} > 0$ such that for all $(\mathbf{B}, \mathbf{\Omega}) \in \mathcal{N}$,

$$\langle \nabla_{\mathbf{B}} \ell(\mathbf{B}, \mathbf{\Omega}) - \nabla_{\mathbf{B}} \ell(\mathbf{B}^\star, \mathbf{\Omega}), \mathbf{B} - \mathbf{B}^\star \rangle \geq \mu_B \|\mathbf{B} - \mathbf{B}^\star\|_F^2,$$
$$\langle \nabla_{\mathbf{\Omega}} \ell(\mathbf{B}, \mathbf{\Omega}) - \nabla_{\mathbf{\Omega}} \ell(\mathbf{B}, \mathbf{\Omega}^\star), \mathbf{\Omega} - \mathbf{\Omega}^\star \rangle \geq \mu_\Omega \|\mathbf{\Omega} - \mathbf{\Omega}^\star\|_F^2,$$

and

$$\|\nabla_{\mathbf{B}} \ell(\mathbf{B}, \mathbf{\Omega}) - \nabla_{\mathbf{B}} \ell(\mathbf{B}, \mathbf{\Omega}^\star)\|_F \leq L_{B\Omega} \|\mathbf{\Omega} - \mathbf{\Omega}^\star\|_F,$$
$$\|\nabla_{\mathbf{\Omega}} \ell(\mathbf{B}, \mathbf{\Omega}) - \nabla_{\mathbf{\Omega}} \ell(\mathbf{B}^\star, \mathbf{\Omega})\|_F \leq L_{\Omega B} \|\mathbf{B} - \mathbf{B}^\star\|_F.$$

Assumption B.4 (Sample-level deviations)**.** The sample loss $\mathcal{L}_n$ satisfies the same RSC and cross-Lipschitz bounds as ℓ, with constants shrunk by at most a factor of $1/2$, and $\|\nabla \mathcal{L}_n(\mathbf{B}^\star, \mathbf{\Omega}^\star)\|_\infty = O\big(\sqrt{(\log p)/n}\big)$.

Under sub-Gaussian tails, Assumption B.4 follows from standard matrix concentration inequalities (e.g. Vershynin, 2018; Wainwright, 2019). Bow-freeness and the eigenvalue margin imply Assumption B.3 by ensuring that the relevant population Hessians are well-conditioned; see Section 2.

Finally, we adopt the usual sparsity and signal-strength assumption:

Assumption B.5 (Sparsity and signal strength)**.** Let $S_B = \mathrm{supp}(\mathbf{B}^\star)$ and $S_\Omega = \mathrm{supp}_{\mathrm{off}}(\mathbf{\Omega}^\star)$ with sizes s_B, s_Ω. For penalties with tuning parameters $\lambda_B, \lambda_\Omega \asymp \sqrt{(\log p)/n}$, assume

$$\min_{(i,j) \in S_B} |B^\star_{ij}| \geq c_B \lambda_B, \qquad \min_{(i,j) \in S_\Omega} |\Omega^\star_{ij}| \geq c_\Omega \lambda_\Omega$$

for constants $c_B, c_\Omega > 1$.

Assumption B.5 is standard in high-dimensional sparse estimation; see, e.g., Bühlmann & Van De Geer (2011); Raskutti et al. (2011).

Proof of Theorem 5.2. Let $e_B^{(t)} = \|\mathbf{B}^{(t)} - \mathbf{B}^\star\|_F$ and $e_\Omega^{(t)} = \|\mathbf{\Omega}^{(t)} - \mathbf{\Omega}^\star\|_F$. The $\mathbf{B}$-update in DECOR minimizes (approximately) the function $\mathbf{B} \mapsto \mathcal{L}_n(\mathbf{B}, \mathbf{\Omega}^{(t)}) + P_B(\mathbf{B}) + \rho h(\mathbf{B})$, keeping $\mathbf{\Omega}^{(t)}$ fixed. By Assumptions B.3–B.4 and the amenable-regularizer analysis of Loh & Wainwright (2015), we obtain the one-step bound

$$e_B^{(t+1)} \;\leq\; \alpha_B e_B^{(t)} + \beta_B e_\Omega^{(t)} + C_B \sqrt{\tfrac{s_B \log p}{n}},$$

for some $\alpha_B < 1$, $\beta_B > 0$, and $C_B > 0$. The first term is a contraction from local strong convexity in $\mathbf{B}$, the second term captures the dependence on $\mathbf{\Omega}^{(t)}$ via the cross-Lipschitz constants, and the last term is statistical error.

Similarly, the $\mathbf{\Omega}$-update minimizes $\mathbf{\Omega} \mapsto \mathcal{L}_n(\mathbf{B}^{(t+1)}, \mathbf{\Omega}) + P_\Omega(\mathbf{\Omega})$ for fixed $\mathbf{B}^{(t+1)}$. Since this subproblem is convex and locally strongly convex near $\mathbf{\Omega}^\star$, standard arguments for ℓ_1- or amenably-regularized covariance estimation yield

$$e_\Omega^{(t+1)} \;\leq\; \alpha_\Omega e_\Omega^{(t)} + \beta_\Omega e_B^{(t+1)} + C_\Omega \sqrt{\tfrac{s_\Omega \log p}{n}},$$

with $\alpha_\Omega < 1$, $\beta_\Omega > 0$, and $C_\Omega > 0$.

Combining the two inequalities and stacking $e^{(t)} = (e_B^{(t)}, e_\Omega^{(t)})^\top$, we obtain a linear recursion

$$e^{(t+1)} \;\leq\; M e^{(t)} + b_n,$$

for a 2×2 matrix M with spectral radius $\rho(M) < 1$ when $\mathcal{N}$ is chosen small enough so that RSC dominates cross-Lipschitz effects, and b_n of order $\delta_n \asymp \sqrt{(s_B \log p)/n} + \sqrt{(s_\Omega \log p)/n}$. Solving the recursion yields $\|e^{(t)}\|_1 \leq \rho(M)^t \|e^{(0)}\|_1 + \|b_n\|_1/(1-\rho(M))$. This proves the claimed contraction and statistical error bound. □

B.3 BOW PROJECTION CONSISTENCY

We expand Lemma 5.4. Recall from Section 5 that for each pair $\{i, j\}$,

$$\widetilde{d}_{ij} = \max\{|\widetilde{B}_{ij}|, |\widetilde{B}_{ji}|\}, \qquad \widetilde{r}_{ij} = \frac{|\widetilde{\Omega}_{ij}|}{\sqrt{\widetilde{\Omega}_{ii}\widetilde{\Omega}_{jj}}},$$

and similarly $d^\star_{ij}$, $r^\star_{ij}$ for the true parameters.

Lemma B.6 (Lipschitz control of r_{ij})**.** *Suppose $\mathbf{\Omega}^\star$ satisfies $\eta\mathbf{I} \preceq \mathbf{\Omega}^\star \preceq M\mathbf{I}$ for some $0 < \eta \leq M$. There exists $L = L(\eta, M)$ such that for any positive definite $\widetilde{\mathbf{\Omega}}$ with $\|\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_\infty \leq 1$,*

$$\left|\widetilde{r}_{ij} - r^\star_{ij}\right| \;\leq\; L\,\|\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_\infty \quad \textit{for all } \{i, j\}.$$

Proof. On the compact set $\mathcal{C} = \{\mathbf{\Omega} : \frac{\eta}{2} \leq \Omega_{kk} \leq 2M,\ \|\mathbf{\Omega}\|_\infty \leq 2M\}$ the map $r_{ij}(\mathbf{\Omega}) = |\Omega_{ij}|/\sqrt{\Omega_{ii}\Omega_{jj}}$ is continuously differentiable. Its partial derivatives are

$$\frac{\partial r_{ij}}{\partial \Omega_{ij}} = \frac{\operatorname{sgn}(\Omega_{ij})}{\sqrt{\Omega_{ii}\Omega_{jj}}}, \quad \frac{\partial r_{ij}}{\partial \Omega_{ii}} = -\frac{|\Omega_{ij}|}{2\Omega_{ii}\sqrt{\Omega_{ii}\Omega_{jj}}},$$

and similarly for Ω_{jj}; each is bounded in absolute value by C/η for a constant C depending on M. Thus the gradient norm $\|\nabla r_{ij}(\mathbf{\Omega})\|_1 \leq L$ for all $\mathbf{\Omega} \in \mathcal{C}$. Since $\mathbf{\Omega}^\star \in \mathcal{C}$ and $\widetilde{\mathbf{\Omega}} \in \mathcal{C}$ whenever $\|\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_\infty \leq 1$, the mean value theorem implies the claimed Lipschitz bound. □

Proof of Lemma 5.4. Let $\delta_{ij} = \widetilde{d}_{ij} - c\widetilde{r}_{ij}$ and $\delta^\star_{ij} = d^\star_{ij} - cr^\star_{ij}$. From the triangle inequality and Lemma B.6,

$$|\delta_{ij} - \delta^\star_{ij}| \;\leq\; |\widetilde{d}_{ij} - d^\star_{ij}| + c|\widetilde{r}_{ij} - r^\star_{ij}| \;\leq\; \varepsilon_B + cL\varepsilon_\Omega.$$

By bow-freeness, for each pair $\{i, j\}$ at most one of $d^\star_{ij}$ and $r^\star_{ij}$ is nonzero. Assumption 5.3 then implies $|\delta^\star_{ij}| \geq \Delta^\star > 0$.

Case 1: True directed edge ($d^\star_{ij} > 0$, $r^\star_{ij} = 0$). Then $\delta^\star_{ij} = d^\star_{ij} \geq \Delta^\star$ and

$$\delta_{ij} \;\geq\; \delta^\star_{ij} - (\varepsilon_B + cL\varepsilon_\Omega) \;>\; \tfrac{1}{2}\Delta^\star > 0,$$

so the rule $\widetilde{d}_{ij} \geq c\widetilde{r}_{ij}$ selects the directed channel. The threshold τ_B ensures that at least one of $|\widetilde{B}_{ij}|, |\widetilde{B}_{ji}|$ exceeds τ_B, so the directed edge is retained.

Case 2: True bidirected edge ($r^\star_{ij} > 0$, $d^\star_{ij} = 0$). Then $\delta^\star_{ij} = -cr^\star_{ij} \leq -\Delta^\star$ and

$$\delta_{ij} \;\leq\; \delta^\star_{ij} + (\varepsilon_B + cL\varepsilon_\Omega) \;<\; -\tfrac{1}{2}\Delta^\star < 0,$$

so the rule selects the bidirected channel; τ_Ω ensures it survives thresholding.

Case 3: No edge ($d^\star_{ij} = r^\star_{ij} = 0$). The error bounds imply $|\widetilde{B}_{ij}|, |\widetilde{B}_{ji}| < \tau_B$ and $|\widetilde{\Omega}_{ij}| < \tau_\Omega$, so no spurious edge is retained.

In all cases, the bow projection recovers the correct edge type (or absence), completing the proof. □

C EXTENDED EXPERIMENTS

Real-world evaluation (Sachs Dataset) We evaluated DECOR on the Sachs protein signaling dataset (Sachs et al., 2005) ($n = 853$, $p = 11$), a standard benchmark for causal discovery. The ground truth is an 18-edge DAG. We compared against several baselines, including standard DAG learners (NOTEARS (Zheng et al., 2018), GOLEM (Ng et al., 2020), GES (Chickering, 2002a)), latent variable methods (DECAMF (Agrawal et al., 2023), DCD (**?**)), and LiNGAM (Shimizu et al., 2006). We also analyzed the effect of our post-hoc reconciliation step. We report results for **DECOR** (the output of the alternating optimization) and **DECOR (bow-strict)** (applying the reconciliation rule to enforce strict bow-freeness).

Table 1: Results on Sachs Data

Method	SHD	Precision	Recall	F1	Edges Pred.
DECOR	**15**	0.616	**0.444**	**0.516**	13
DECOR (bow-strict)	**15**	**0.714**	0.278	0.400	7
GES	16	0.571	0.444	0.500	14
DECOR_GL	18	0.500	0.167	0.250	6
DECOR_GL (bow-strict)	18	0.500	0.111	0.182	4
LiNGAM	19	0.455	0.222	0.348	11
DECAMF (LIN r1)	19	0.444	0.222	0.296	9
DCD	19	0.333	0.056	0.095	3
GOLEM	20	0.375	0.167	0.231	10
NOTEARS	22	0.167	0.056	0.083	6
AdaScore	25	0.333	0.389	0.359	21

Key Findings: DECOR achieves the lowest SHD (15) and highest F1 score (0.516), outperforming both classic methods (LiNGAM, GES) and recent latent-variable baselines (DCD, AdaScore). The `bow-strict` reconciliation acts as a rigorous filter: it significantly improves **Precision** (from 0.616 to 0.714, the highest among all methods) by enforcing the mutual exclusivity of directed and bidirected edges. However, this comes at the cost of **Recall** (dropping from 0.444 to 0.278) in this specific dataset, as weaker causal signals are conservatively assigned to the noise covariance. **DECOR vs. DECOR_GL:** The covariance-based formulation (DECOR) is more liberal, recovering more edges (13 predicted) compared to the precision-based formulation (DECOR_GL, 6 predicted), which appears over-penalized on this dataset.