Regularization Paths for Continuous DAG Learning

Diyang Li
Cornell University
diyang01@cs.cornell.edu

Fei Wang
Cornell University
few2001@med.cornell.edu

Kyra Gan
Cornell University
kyragan@cornell.edu

Abstract

Continuous DAG learning formulates combinatorial structure learning as differentiable optimization over weighted adjacency matrices. In these methods, the estimated adjacency matrix depends critically on a sparsity-controlling regularization parameter, whose magnitude governs the recovered graph skeleton. While extensive empirical evidence underscores the centrality of this parameter, existing practice relies on solving a sequence of independent optimization problems over a grid of tuning parameters, which is computationally wasteful, statistically unstable, and provides only a fragmented view of how the adjacency matrix evolves. In this work, we study the regularization path of stationary solutions in continuous DAG learning and propose `DAG-flow`, an exact path-following framework that compactly encodes the entire family of estimators. We show that for prominent formulations including NOTEARS, GOLEM, and DAGMA, stationary solutions evolve according to a piecewise-smooth matrix-valued dynamical system once the active support is fixed. Our analysis derives the explicit systems governing these branches, identifies the events at which the path changes regime, and clarifies how nonsmooth sparsity and nonconvex acyclicity interact along the path. Unlike grid-based search, `DAG-flow` exposes the geometry of model variation across regularization levels, which enables the construction of structured candidate graph sets and provides direct insight into edge-level stability. Across synthetic and real benchmarks, our `DAG-flow` outperforms grid-search baselines at a fraction of the compute and consistently improves structural stability.

1 Introduction

Continuous DAG Learning. Recovering the structure of a directed acyclic graph (DAG) from observational data is a fundamental problem in causal discovery and probabilistic graphical modeling, with applications spanning machine learning, economics, and neuroscience [Pearl, 2009, Koller and Friedman, 2009, Morgan and Winship, 2014]. Classical score-based and constraint-based approaches treat this as a combinatorial search over the space of feasible DAGs, an NP-hard problem whose super-exponential cost in the number of nodes severely limits scalability [Chickering et al., 2004]. Continuous DAG learning [Zheng et al., 2018] addresses this challenge by representing a graph through a weighted adjacency matrix and replacing discrete graph search with a continuous optimization problem over real matrices. Since the introduction of smooth acyclicity characterizations, methods such as NOTEARS [Zheng et al., 2018], GOLEM [Ng et al., 2020], and DAGMA [Bello et al., 2022] have shown that DAG structure learning can be formulated as a continuous optimization problem involving a data-fitting score, an acyclicity constraint or soft penalty, and a sparsity-inducing regularizer. The recovered graph is then read off from the support of the resulting weighted adjacency matrix. This perspective has enabled the use of well-studied gradient-based algorithms and has become a dominant framework for scalable DAG estimation [Zheng et al., 2020].

In most continuous DAG learning methods, an ℓ_1-type penalty term is imposed on the weighted adjacency matrix. The associated regularization parameter controls the magnitude of edge shrinkage

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and therefore determines the density of the estimated graph skeleton [Ng et al., 2020]. Extensive empirical evidence across the literature confirms that the recovered structure is highly sensitive to this tuning parameter [Zheng et al., 2018, Ng et al., 2020, Xu et al., 2022, Saboksayr et al., 2024], with small values yielding dense graphs that overfit spurious dependencies and overly large values shrinking genuine edges and corrupting the acyclic backbone. Despite the universally acknowledged importance, the current practice is unsatisfying: one typically handles this sensitivity by solving the DAG learning problem repeatedly over a grid of regularization parameters, and model selection is then performed by comparing the resulting graphs using structural metrics or downstream performance [Lachapelle et al., 2020, Bello et al., 2022]. This approach is simple and widely used, but it is also inherently fragmented and statistically unstable [Park and Hastie, 2007, Friedman et al., 2010]. It treats nearby regularization levels as unrelated optimization problems, discards the strong continuity between neighboring solutions, and often requires substantial computation because each grid point may involve a difficult nonconvex optimization problem [Wei et al., 2020, Ng et al., 2022].

Regularization Path. Within this paradigm, a more principled view is to study the *regularization path* (see Fig. 1), i.e., the entire trajectory of estimators traced out as the regularization parameter varies continuously [Efron et al., 2004]. Rather than computing a finite collection of independent estimates, the path provides a compact representation of how the weighted adjacency matrix evolves across regularization levels. Tracing the path yields, at a small fraction of the cost of a sufficiently fine grid, a structured candidate set of estimators in place of a single-point estimate, precisely the object needed for downstream model selection [Li et al., 2024], ensembling [Khaire and Dhanalakshmi, 2022], or uncertainty quantification [Meinshausen and Bühlmann, 2010]. Meanwhile, it provides direct insight into edge-level stability, since edges that survive along broad portions of the path are more robust than those that appear only under narrow tuning regimes [Manjunath et al., 2025]. From a computational standpoint, solution path methods can maintain a more coherent trajectory of estimates than discrete retraining, which avoids potential numerical instability concerns [Rosset and Zhu, 2007]. These observations motivate a practical path-following treatment of continuous DAG learning.

Technical Challenges. Extending path-following ideas to continuous DAG learning is technically delicate. Most existing (exact) solution path frameworks were developed for the classical unconstrained empirical risk minimization [Friedman et al., 2023], while comparatively fewer path-following results for constrained optimization rely on quadratic or convex constraints [Zhou and Lange, 2015], under which the feasible set is a polyhedron and the path is governed by a relatively regular KKT system. The continuous DAG learning violates every one of these settings simultaneously. The acyclicity function is a global non-convex constraint: the feasible DAG set is not a single smooth manifold, but a union of strata associated with different topological orders. Moreover, in constrained formulations such as NOTEARS and DAGMA, the acyclicity characterization is first-order analytically degenerate on the DAG space [Wei et al., 2020], so a naive application of the implicit function theorem on the KKT conditions does not produce a valid homotopy direction. One must instead work locally on a fixed acyclic stratum, freeze the relevant topological order, and derive the stationary dynamics only within that regime. This difficulty is further compounded by the non-smooth elementwise ℓ_1 penalty in the primal objective. Finally, the object studied in DAG learning is intrinsically a matrix, and the dynamical system governing its evolution is matrix-valued, with structure that couples columns through the acyclicity term and rows through the score, so the path equations cannot be reduced to a vectorized form without obscuring the very geometry one wishes to expose.

Our Contributions. We develop `DAG-flow`, an exact path-following framework for the three major continuous DAG learning paradigms in the field, i.e., NOTEARS, GOLEM, and DAGMA. Our first contribution is a path-level characterization of stationary solutions in these models. For the exact constrained formulations NOTEARS and DAGMA, we identify the zero-solution threshold and show that, once the active support and the local acyclic regime are fixed, the regularization path is piecewise affine: active edge weights follow explicit linear matrix flows, while inactive entries remain zero until a transition point is reached. For GOLEM, where acyclicity is incorporated through a smooth penalty and the objective contains a log-determinant term, we study the two natural coordinate paths obtained by varying either the sparsity-controlling hyperparameter or the acyclicity-penalty hyperparameter, and show that both paths are piecewise smooth with explicit matrix-valued flow representations. Our second contribution is an event characterization for continuous DAG paths. We identify the regime changes that terminate a regular path segment, including edge deletion, insertion, topological transitions for exact DAG constraints, and the additional domain and turning events that arise in GOLEM. These events provide the update rules that allow `DAG-flow` to move from one regular

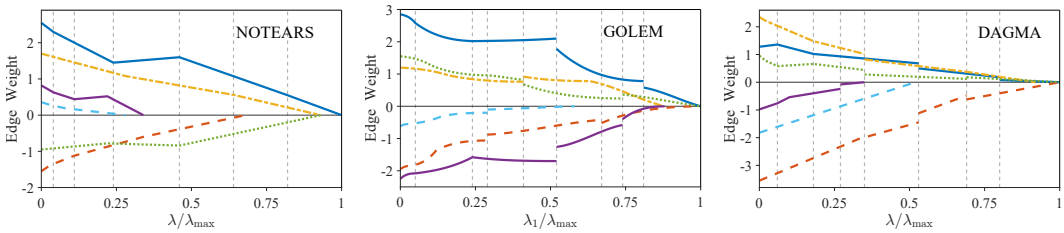


Figure 1. Visualization of the regularization path on example synthetic data with $d = 4$ nodes. Each piecewise-smooth curve represents the solution path of one entry, and each vertical line marks a transition point (kink).

regime to the next efficiently. Our third contribution is a finite convergence result: under generic data conditions, the stationary solution graph decomposes into finitely many analytic path segments rather than an unstructured continuum of unrelated solutions. Lastly, we instantiate the `DAG-flow` as an exact path-following algorithm and numerically evaluate it on synthetic and real benchmarks, where it consistently outperforms grid-search baselines at a fraction of the compute.

Overall, we position this work as a foundational *algorithmic interface* rather than an end-point solution in DAG learning. By exposing the entire regularization path geometry in closed analytic form, it opens a direct pathway for future downstream applications and methodological extensions, ranging from model selection and post-selection inference to sensitivity analysis and ensembling, that are otherwise inaccessible under conventional grid-based practice.

1.1 Related Work

Solution path in DAG learning. The idea of tracing a path in continuous DAG learning is not entirely new, but existing works pursue different goals from our `DAG-flow`. Several path-like procedures are designed to vary a penalty or Lagrangian parameter in order to approximately enforce acyclicity, rather than to characterize the sparsity regularization path of stationary graph estimators. Wei et al. [2020] study the behavior of continuous DAG optimization more closely and develop a LARS-type procedure for edge insertions or deletions. Bello et al. [2022] introduce a penalty path in their central-path interpretation, but the resulting trajectory is used as a warm-start mechanism rather than analyzed as a structured object. More recently, Deng et al. [2023] analyze global penalty paths in the bivariate setting, while Negri et al. [2023] introduce matrix-flow ideas for related Gaussian graphical model problems, but the flow learns an amortized surrogate rather than tracking exact stationary solutions. In addition, Zhang et al. [2025] adopt a path-following perspective on the Lagrangian multiplier in differentiable DAG learning, varying the multiplier of the analytic acyclicity constraint.

Regularization path for structure learning. A second line of work does target the sparsity regularization parameter in DAG-structured estimation, but uniformly resorts to grid-based re-optimization or discrete heuristic search to approximate the path, including Schmidt et al. [2007], Vidaurre et al. [2010], Fu and Zhou [2013], Aragam and Zhou [2015], Li et al. [2018], Gu et al. [2019], Aragam et al. [2019], Li et al. [2024], Manjunath et al. [2025]. While this collective effort confirms the practical importance of the regularization path, it stops at sampling the path on a finite grid and discards both the exactness and the analytic structure that a true path-following framework provides.

2 Preliminaries

DAG Models. We observe n i.i.d. samples of d variables, stored row-wise in $\mathbf{X} \in \mathbb{R}^{n\times d}$, with $X_i \in \mathbb{R}^d$ denoting the i-th sample. We focus on linear Structural Equation Models (SEMs) [Bollen, 2014], which in matrix form reads $\mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{N}$, where $\mathbf{W} = [\mathbf{W}_1 \mid \cdots \mid \mathbf{W}_d] \in \mathbb{R}^{d\times d}$ is the weighted adjacency matrix and $\mathbf{N}$ is the noise matrix. For a matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, $\mathbf{W}_{ij}$ denotes its (i, j)-th entry. A nonzero entry $\mathbf{W}_{ij}$ represents the directed edge $i \to j$. A function is said to be C^∞ if it is infinitely differentiable.

Regularization Path. Following standard practice, we work at the level of *stationary points* of the first-order conditions rather than local or global optima: the primal objectives of NOTEARS, GOLEM, and DAGMA are all highly non-convex, and stationary solutions are widely accepted in the community as the operative solution concept [Zheng et al., 2018, Ng et al., 2020]. Let $\lambda > 0$ denote

the regularization parameter, and $\hat{\mathbf{W}}$ denote a stationary point. Given an initial $\hat{\mathbf{W}}(\lambda_0)$ at $\lambda = \lambda_0$, our goal is to track a continuous branch $\lambda \mapsto \hat{\mathbf{W}}(\lambda)$ over the entire feasible interval of λ; Figure 1 illustrates a simple example. Throughout the paper we restrict attention to NOTEARS, GOLEM, and DAGMA, and impose Assumption 2.1, which holds almost surely under generic data-generating processes for $d \geq 2$. To lighten notation we write $\mathbf{W}(\lambda)$ for $\hat{\mathbf{W}}(\lambda)$ when no ambiguity arises.

Assumption 2.1. For each of the three objectives, there exists at least one solution at every $\lambda > 0$, and on any compact interval $[a, b] \subset (0, \infty)$ the tracked branch has $\sup_{\lambda \in [a,b]} \|\mathbf{W}(\lambda)\| < \infty$.

Under Assumption 2.1, our subsequent analysis will show that the path $\mathbf{W}(\lambda)$ is piecewise smooth in λ, with non-smoothness confined to an isolated set of switch points at which one of two structural events occurs. We make this precise via the following definition.

Definition 2.2 (Transition Point)**.** Fix $\lambda_0 > 0$ and let $\mathbf{W}(\lambda_0)$ be a stationary point of a continuous DAG learning problem. We call $(\mathbf{W}(\lambda_0), \lambda_0)$ a *transition point*, also referred to as a *kink* or *critical point* of the regularization path, if the stationarity system fails to remain in a single regular analytic regime at $(\mathbf{W}(\lambda_0), \lambda_0)$.

Transition points occur when the local stationary system changes its functional form [Tibshirani and Taylor, 2011]. Equivalently, after freezing the sign pattern $\operatorname{sign}(\mathbf{W}(\lambda))$ and a topological order of the DAG encoded by $h(\mathbf{W}(\lambda_0)) = 0$, at least one of the following occurs: **(i)** a *sign transition*, in which the local representation of $\|\mathbf{W}\|_1$ becomes unstable, i.e., a zero entry that is locally feasible under the frozen topological order satisfies its inactive subgradient condition with equality; or **(ii)** a *topological transition*, in which the active acyclicity manifold $\{h(\mathbf{W}) = 0\}$ admits no local parameterization by the same topological order, so the path leaves the current acyclic stratum. Note that, unlike in the NOTEARS (or DAGMA) hard-constrained case, no separate topological-order transition arises for GOLEM, where h enters as an analytic soft penalty rather than a hard constraint.

3 Regularization Path for NOTEARS

We first instantiate our path-following analysis on NOTEARS [Zheng et al., 2018]. At a glance, we will show that the path lives on a bounded λ-interval (Theorem 3.1), is piecewise real-analytic (Theorem 3.2), is governed by an explicit linear system on each segment (Theorem 3.3), and has closed-form transition points (Theorem 3.4). Concretely, the NOTEARS problem solves

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times d}} \frac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_1 \quad \text{s.t.} \quad h(\mathbf{W}) = 0, \tag{1}$$

where $h(\mathbf{W}) := \operatorname{tr}(e^{\mathbf{W} \odot \mathbf{W}}) - d$ is the smooth acyclicity characterization, $\|\cdot\|_1$ denotes the entrywise ℓ_1-norm, and $\lambda > 0$ controls sparsity of the recovered graph. Our goal is to characterize the entire stationary branch $\lambda \mapsto \mathbf{W}(\lambda)$ rather than solve (1) at a grid of λ values. The regularization path is meaningful only on a bounded λ-interval, since for sufficiently large λ the trivial solution dominates. The right endpoint admits a closed form.

Theorem 3.1 (Zero-solution Threshold)**.** *Consider the NOTEARS problem (1), and let*

$$\lambda_{\max} = \left\| n^{-1} \mathbf{X}^\top \mathbf{X} - \operatorname{diag}\left(n^{-1} \mathbf{X}^\top \mathbf{X}\right) \right\|_{\max},$$

where $\|\cdot\|_{\max}$ denotes the entrywise absolute maximum. Then $\mathbf{W} = \mathbf{0}$ is a global minimizer of the NOTEARS objective for all $\lambda \geq \lambda_{\max}$, and is the unique one whenever $\lambda > \lambda_{\max}$ and $\lambda_{\max} > 0$.

Theorem 3.1 restricts the path domain to $\lambda \in (0, \lambda_{\max}]$ and supplies a canonical right endpoint. We next investigate the local structure of the path. At an exact DAG, we know $\nabla h(\mathbf{W}) = 0$, and the hard acyclicity constraint does not contribute an ordinary nonzero multiplier force along a feasible stratum. Instead, after fixing a compatible topological order, the constraint locally removes the entries that would leave the current acyclic stratum. This is the key distinction between NOTEARS and an ordinary smooth equality-constrained problem.

Theorem 3.2 (Local Smoothness)**.** *Fix $0 < \lambda_0 < \lambda_{\max}$, and let $\mathbf{W}(\lambda_0)$ be a stationary point of (1). Suppose that $(\mathbf{W}(\lambda_0), \lambda_0)$ is not a transition point. Then there exist $\varepsilon > 0$ and a real-analytic map*

$$\lambda \mapsto \mathbf{W}(\lambda), \quad \lambda \in (\lambda_0 - \varepsilon, \lambda_0 + \varepsilon),$$

such that $\mathbf{W}(\lambda)$ is a stationary point of (1) for every such λ. Moreover, the local regularization path is C^∞ smooth and admits an analytic continuation up to the nearest transition point.

Away from transition points (Definition 2.2), Theorem 3.2 shows that the stationary branch is not merely continuous but real-analytic. We now derive the dynamical system governing the active entries on a smooth segment. Let $\mathcal{A} = \{(i,j) : \mathbf{W}_{ij}(\lambda) \neq 0\}$ denote the **active set** on the current segment of the regularization path, and for each node j let $\mathcal{A}_j = \{i : (i,j) \in \mathcal{A}\}$ be its active parent set, with $\mathbf{X}_{:,\mathcal{A}_j}$ being the corresponding data columns; if $\mathcal{A}_j = \varnothing$ the related equation is void.

Theorem 3.3 (Matrix flow)**.** *Consider* (1)*. Under the hypotheses of Theorem 3.2, the active entries of the local regularization path satisfy the first-order linear differential system*

$$\left(\frac{1}{n}\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}\right)\frac{d\mathbf{W}_{\mathcal{A}_j,j}}{d\lambda} = -\operatorname{sign}\left(\mathbf{W}_{\mathcal{A}_j,j}\right), \qquad j = 1,\dots,d, \tag{2}$$

and every inactive entry remains zero on the same local regime until a transition point is reached.

Two observations are in order. First, the system (2) *decouples across columns*: the dynamics of the parents of node j depend only on the column-j Gram block $\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}$, so the global d^2-dimensional flow factors into d independent linear systems of size $|\mathcal{A}_j|$. This factorization is what makes column-wise integration tractable even at large d. Second, the right-hand side is the constant sign vector $\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j})$, which is frozen on each segment by definition of the regular regime; consequently the path is in fact *piecewise affine* in λ on each segment, and the segment can be advanced in closed form. A path-tracking algorithm also identifies the smallest $\lambda > \lambda_0$ at which the current regime breaks. As λ increases, the nearest non-smooth transition occurs at a closed-form value λ_{next} , as follows.

Theorem 3.4 (Nearest Transition)**.** *Consider* (1)*. Suppose that $\mathbf{W}(\lambda_0)$ is a stationary point. Assume that we follow the stationary branch in the increasing-λ direction, and let $\dot{\mathbf{W}}_{ij} = \frac{d\mathbf{W}_{ij}}{d\lambda}$. For every inactive pair $(i,j) \notin \mathcal{A}$ whose insertion $i \to j$ preserves acyclicity, define*

$$g_{ij} = \frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda_0) - \mathbf{X}_{:,j}\right), \quad \dot{g}_{ij} = \frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\dot{\mathbf{W}}_{kj}\right).$$

Define the active-drop candidate and inactive-entry candidate as

$$\lambda_{\text{drop}} = \min_{\substack{(i,j)\in\mathcal{A},\,\dot{\mathbf{W}}_{ij}\neq 0\\ \lambda_0<\lambda_0-\mathbf{W}_{ij}(\lambda_0)/\dot{\mathbf{W}}_{ij}\le\lambda_{\max}}}\left\{\lambda_0 - \frac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}}\right\}, \quad \lambda_{\text{enter}} = \min_{(i,j)\notin\mathcal{A}}\left\{\max\left\{\frac{g_{ij}-\lambda_0\dot{g}_{ij}}{1-\dot{g}_{ij}}, \frac{\lambda_0\dot{g}_{ij}-g_{ij}}{1+\dot{g}_{ij}}\right\}\right\}.$$

Then the nearest transition point in the increasing-λ direction is $\lambda_{\text{next}} = \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}$.

If $1 \pm \dot{g}_{ij} = 0$ the corresponding candidate in $\max\{\cdot,\cdot\}$ is omitted, and if both are omitted we set $\lambda_{\text{enter}} = +\infty$. Combined with Theorem 3.3, Theorem 3.4 reduces full-path computation for NOTEARS to alternating piecewise advances with no inner optimization loop. At λ_{next}, the algorithm updates the support or the acyclic stratum and restarts the same matrix-flow construction.

4 Regularization Path for GOLEM

We next study GOLEM [Ng et al., 2020], whose optimization objective is

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times d}} \frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_1\|\mathbf{W}\|_1 + \lambda_2 h(\mathbf{W}), \tag{3}$$

where $\lambda_1 > 0$ controls sparsity and $\lambda_2 > 0$ controls the soft acyclicity penalty. We present the equal-variance (EV) case; the non-equal-variance case follows analogously. Unlike NOTEARS, GOLEM uses $h(\mathbf{W})$ as a smooth penalty rather than a hard constraint. Hence there is no separate topological-order transition since acyclicity affects the path through the smooth gradient of h. Note that GOLEM has *two* regularization knobs, so the stationary set forms a two-parameter surface $(\lambda_1,\lambda_2) \mapsto \mathbf{W}(\lambda_1,\lambda_2)$ rather than a one-parameter curve. We track this surface along its two coordinate paths: the λ_1-path (sparsity sweep at fixed λ_2) and the λ_2-path (acyclicity sweep at fixed λ_1). Let the admissible domain be given by $\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 > 0$ and $\det(\mathbf{I}-\mathbf{W}) \neq 0$ [Ng et al., 2020]. For fixed λ_2, the relevant λ_1-range is determined by the zero-solution threshold.

Theorem 4.1 (Zero-solution Threshold)**.** *Consider the GOLEM problem* (3)*, and let*

$$\lambda_{\max} = \max\left\{0, \sup_{\substack{\mathbf{W}\neq\mathbf{0}\\ \det(\mathbf{I}-\mathbf{W})\neq 0\\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F>0}} \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right\}.$$

Then $\mathbf{W} = \mathbf{0}$ is a global minimizer of the GOLEM objective for all $\lambda_1 \geq \lambda_{\max}$, and is the unique one whenever $\lambda_1 > \lambda_{\max}$. A necessary and sufficient condition for the $\lambda_{\max}$ to be finite is $\mathbf{X}^\top\mathbf{X} \succ 0$.

The finiteness condition $\mathbf{X}^\top\mathbf{X} \succ 0$ holds generically when $n \geq d$. In the high-dimensional regime $n < d$, GOLEM may not admit a finite zero-solution threshold. We then show that the stationary surface is real-analytic in both arguments simultaneously, not merely along each coordinate.

Theorem 4.2 (Local Smoothness for GOLEM)**.** *Fix $(\lambda_1^0, \lambda_2^0)$ in the interior of the admissible region $(0, \lambda_{\max}) \times (0, +\infty)$, and let $\mathbf{W}(\lambda_1^0, \lambda_2^0)$ be a stationary point of (3). If $\big(\mathbf{W}(\lambda_1^0, \lambda_2^0), \lambda_1^0, \lambda_2^0\big)$ is not a transition point of the GOLEM solution surface, then there exists $\varepsilon > 0$ and a real-analytic map*

$$(\lambda_1, \lambda_2) \mapsto \mathbf{W}(\lambda_1, \lambda_2), \qquad (\lambda_1, \lambda_2) \in \mathbb{B}_\varepsilon\big((\lambda_1^0, \lambda_2^0)\big),$$

where $\mathbb{B}_\varepsilon\big((\lambda_1^0, \lambda_2^0)\big) := \big\{(\lambda_1, \lambda_2) : (\lambda_1 - \lambda_1^0)^2 + (\lambda_2 - \lambda_2^0)^2 < \varepsilon^2\big\}$ such that $\mathbf{W}(\lambda_1, \lambda_2)$ is a stationary point of (3). The coordinate paths $\lambda_1 \mapsto \mathbf{W}(\lambda_1, \lambda_2^0)$, $\lambda_2 \mapsto \mathbf{W}(\lambda_1^0, \lambda_2)$ are C^∞ smooth in a neighborhood of λ_1^0 and λ_2^0, respectively. Finally, the local surface admits analytic continuation along any connected transition-free subset of the parameter region, up to the nearest transition locus.

On a transition-free segment of either coordinate path, the active entries evolve according to a smooth linear ordinary differential equation (ODE) in matrix form. Inactive entries remain frozen at zero, as in NOTEARS, so it suffices to characterize the active velocity matrix.

Theorem 4.3 (Matrix Flow of GOLEM)**.** *Fix a transition-free segment of a stationary solution path of (3) and let $\mathbf{V}^{(k)}_{\mathcal{A}_j, j} := d\mathbf{W}_{\mathcal{A}_j, j}/d\lambda_k$ for $k \in \{1, 2\}$. Inactive entries satisfy $d\mathbf{W}_{\mathcal{A}^c}/d\lambda_k = \mathbf{0}$. The active velocities are determined columnwise, for $j = 1, \ldots, d$, by the linear system $\mathcal{L}_{\lambda_2}[\mathbf{V}^{(k)}]_{\mathcal{A}_j, j} = \mathbf{r}^{(k)}_{\mathcal{A}_j, j}$, where the operator $\mathcal{L}_{\lambda_2}$ acts on a candidate velocity matrix $\mathbf{V}$ via*

$$\mathcal{L}_{\lambda_2}[\mathbf{V}]_{\mathcal{A}_j, j} = \frac{d\,\mathbf{X}^\top_{:,\mathcal{A}_j}\mathbf{X}_{:,\mathcal{A}_j}\mathbf{V}_{\mathcal{A}_j, j}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}^\top_{:,\mathcal{A}_j}(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^{d} \big[\mathbf{X}^\top_{:,\mathcal{A}_r}(\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})\big]^\top \mathbf{V}_{\mathcal{A}_r, r}$$

$$+ \big[(\mathbf{I} - \mathbf{W})^{-\top}\mathbf{V}^\top(\mathbf{I} - \mathbf{W})^{-\top}\big]_{\mathcal{A}_j, j} + 2\lambda_2 \Big\{ \mathbf{V}_{\mathcal{A}_j, j} \odot \big[(e^{\mathbf{W}\odot\mathbf{W}})^\top\big]_{\mathcal{A}_j, j} + \mathbf{W}_{\mathcal{A}_j, j} \odot \big[\mathcal{D}h(\mathbf{W})[\mathbf{V}]^\top\big]_{\mathcal{A}_j, j} \Big\},$$

with $\mathcal{D}h(\mathbf{W})[\mathbf{V}] = \int_0^1 e^{(1-t)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W} \odot \mathbf{V})\, e^{t(\mathbf{W}\odot\mathbf{W})}\, dt$ the Fréchet derivative of h at $\mathbf{W}$. The right-hand sides are

$$\mathbf{r}^{(1)}_{\mathcal{A}_j, j} = -\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j, j}), \qquad \mathbf{r}^{(2)}_{\mathcal{A}_j, j} = -2\,\mathbf{W}_{\mathcal{A}_j, j} \odot \big[(e^{\mathbf{W}\odot\mathbf{W}})^\top\big]_{\mathcal{A}_j, j}, \quad j = 1, \ldots, d.$$

Three structural points distinguish Theorem 4.3 from the NOTEARS flow (2). First, the two coordinate flows share the *same* linear operator $\mathcal{L}_{\lambda_2}$; only the forcing term differs, so a single factorization of $\mathcal{L}_{\lambda_2}$ supports both sweeps. Second, the second and third terms of $\mathcal{L}_{\lambda_2}$ couple the columns of $\mathbf{V}$ through the residual outer product and the inverse $(\mathbf{I} - \mathbf{W})^{-\top}$: the columnwise decoupling enjoyed by NOTEARS is lost, and a global $|\mathcal{A}|$-dimensional system must be solved per step. Third, although $\mathcal{L}_{\lambda_2}$ is no longer constant in $\mathbf{W}$, the segment ODE is *genuinely linear* in the velocity, so each predictor step reduces to one linear solve rather than a nonlinear root-finding subproblem.

Theorem 4.4 (Nearest Transition for GOLEM)**.** *Consider (3). With the convention $\inf \varnothing = +\infty$, the nearest forward transition point along the λ_1-path at $\lambda_2 = \lambda_2^0$ is*

$$\lambda_{1,\text{next}} = \min\{\lambda_{1,\text{drop}},\, \lambda_{1,\text{enter}},\, \lambda_{1,\text{dom}},\, \lambda_{1,\text{turn}},\, \lambda_{\max}\},$$

and analogously $\lambda_{2,\text{next}} = \min\{\lambda_{2,\text{drop}},\, \lambda_{2,\text{enter}},\, \lambda_{2,\text{dom}},\, \lambda_{2,\text{turn}}\}$ along the λ_2-path at $\lambda_1 = \lambda_1^0$, where each candidate is defined in Appendix A.1.

If the minimum is attained by $\lambda_{1,\text{enter}}$, then all inactive entries satisfying

$$\left| \frac{d\,[\mathbf{X}^\top\mathbf{X}(\mathbf{W} - \mathbf{I})]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I} - \mathbf{W})^{-\top}]_{ij} \right| := \big|\phi_{ij}(\mathbf{W})\big| = \lambda_{1,\text{next}}, \qquad i \notin \mathcal{A}_j,$$

are inserted into the active set with entering sign $\operatorname{sign}(\mathbf{W}_{ij}) = -\operatorname{sign}(\phi_{ij})$. Combined with Theorem 4.3, Theorem 4.4 yields a complete forward-tracking scheme on either coordinate path: between events, integrate the segment ODE; at events, update $\mathcal{A}$ and re-factorize $\mathcal{L}_{\lambda_2}$.

5 Regularization Path for DAGMA

We finally consider the log-determinant acyclicity characterization used in DAGMA [Bello et al., 2022]. For a fixed $s > 0$, define

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times d}} \quad \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1 \quad \text{s.t.} \quad h^s_{\text{ldet}}(\mathbf{W}) = 0, \tag{4}$$

where $h^s_{\text{ldet}}(\mathbf{W}) := -\log\det(s\mathbf{I} - \mathbf{W}\odot\mathbf{W}) + d\log s$. Like the NOTEARS constraint, h^s_{ldet} is an exact acyclicity certificate, which has the same feasible set as the exact NOTEARS formulation; unlike it, h^s_{ldet} is real-analytic only on $\{\mathbf{W} : s\mathbf{I} - \mathbf{W}\odot\mathbf{W} \succ 0\}$ and exhibits markedly better-conditioned gradients in practice. Hence the DAGMA path inherits the structural sketch of the NOTEARS path in Section 3; we therefore present the results in compressed form.

Theorem 5.1 (Zero-solution Threshold, DAGMA)**.** *Consider the DAGMA problem* (4)*, and let* $\lambda_{\max} = \left\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \operatorname{diag}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\right\|_{\max}$*. Then* $\mathbf{W} = \mathbf{0}$ *is a global minimizer of the DAGMA objective for all* $\lambda \geq \lambda_{\max}$*, and is the unique one whenever* $\lambda > \lambda_{\max}$ *and* $\lambda_{\max} > 0$*.*

Intuitively, the threshold in Theorem 5.1 coincides with that of Theorem 3.1 because at $\mathbf{W} = \mathbf{0}$ both acyclicity scores satisfy $h(\mathbf{0}) = h^s_{\text{ldet}}(\mathbf{0}) = 0$ with vanishing gradient.

Corollary 5.2 (Local Smoothness, DAGMA)**.** *Theorem 3.2 holds verbatim for* (4) *on the effective domain* $\{\mathbf{W} : s\mathbf{I} - \mathbf{W}\odot\mathbf{W} \succ 0\}$*.*

Corollary 5.3 (Matrix Flow, DAGMA)**.** *Theorem 3.3 holds verbatim for* (4)*: the active entries of the stationary path satisfy the column-decoupled affine system* (2)*, and inactive entries remain zero.*

The KKT system of (4) differs from that of (1) only through the multiplier of the acyclicity constraint, but on any segment whose active topological order is preserved the constraint contribution decouples from the active sub-system that drives $d\mathbf{W}/d\lambda$. Consequently, both the local smoothness statement and the column-wise matrix flow carry over. In particular, the path remains *piecewise affine* in λ.

Corollary 5.4 (Transition Boundary, DAGMA)**.** *For the exact constrained path of* (4)*, Theorem 3.4 applies without modification:* $\lambda_{\text{next}} = \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}$*.*

Corollary 5.4 should be read with care: it concerns the *exact* constrained stationary path; away from $\{h^s_{\text{ldet}} = 0\}$, the two scores disagree, and the equivalence of Corollary 5.4 does *not* extend to either an augmented-Lagrangian path, or the DAGMA central path tracked via s. Combined with Corollary 5.3, the present Corollary 5.4 suffices for full-path computation under the formulation (4).

6 Algorithmic Overview

Tracking scheme. The preceding sections show that stationary regularization paths are locally analytic between transition points, which leads to a simple path-following principle. For NOTEARS and exact DAGMA, the resulting dynamics reduce to the linear systems in Theorem 3.3 and Corollary 5.3. Since the coefficients are constant within a regime, each segment is affine in λ, and the next transition can be obtained by solving the explicit entry and drop conditions. For GOLEM, the stationary equations are smooth but nonlinear in $\mathbf{W}$. Thus, `DAG-flow` solves the active linearized system in Theorem 4.3 at each step and numerically integrates the resulting initial-value problem along either the λ_1-path or the λ_2-path, using any mature off-the-shelf ODE solver [Butcher, 2016].

At a transition point, the current analytic regime is no longer valid. If an active coefficient reaches zero, the edge is removed; if an inactive coefficient satisfies its entry condition, the edge is inserted with the corresponding one-sided sign. For exact NOTEARS and DAGMA, the `DAG-flow` also checks whether the current acyclic stratum remains valid. For GOLEM, domain and turning events are additionally monitored. When the path is continuous through the event, `DAG-flow` simply updates the regime and continues. If the stationary branch jumps, the last valid solution is used as a warm-start to recover the next stationary point and restart the continuation. Our `DAG-flow` replaces a grid of non-convex optimizations with one continuation procedure, yielding substantial gains in efficiency.

Finite Convergence. A natural concern is whether this event-driven loop can run indefinitely. The next theorem rules this out: under generic data, the path decomposes into a finite number of analytic pieces. Theorem 6.1 does not assert that a single initialization visits every stationary branch; rather,

it guarantees that the exact stationary graph is stratified into finitely many regimes. This finiteness is what makes the path representation substantially more compact than grid search: each segment encodes a continuum of regularization values, while grid search observes only isolated points.

Theorem 6.1 (Finite Analytic Stratification)**.** *Assume that* $\mathrm{vec}(\mathbf{X})$ *has a joint density with respect to Lebesgue measure on* $\mathbb{R}^{nd}$*. Then, almost surely, the number of transition points (TPs) on the exact stationary solution graph of each of NOTEARS* (1) *and DAGMA* (4) *satisfies*

$$\#\left\{\text{TPs in } 0 < \lambda < +\infty\right\} \le 1 + d! \left[\frac{1}{2}\sum_{k=0}^{d-1}\left(\sum_{r=0}^{\min(n,k)} 2^r \binom{k}{r} - 1\right) - 1\right].$$

For GOLEM, assume that the stationary branches are contained in $\{\mathbf{W} : \|\mathbf{W}\|_F \le M\}$ *inside the effective domain. The maximum* λ_2 *under user consideration is* $\lambda_2^{\max}$*. Fix* $\lambda_2 = \lambda_2^0$ *for* λ_1*-path, and fix* $\lambda_1 = \lambda_1^0$ *for* λ_2*-path. Then, with probability one, for* $q \in \{1, 2\}$*, the number of TPs of* (3) *satisfies*

$$\#\left\{\text{TPs along } \lambda_q\text{-path}\right\} \le \sum_{(r,c)\in\{1,\dots,d\}^2} \sum_{\sigma=\pm 1} \sum_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}} \sum_{s\in\{\pm1\}^S} \#Z^{(q)}_{(r,c),\sigma,S,s} + \sum_{\varnothing\neq S\subseteq\{1,\dots,d\}^2} \sum_{s\in\{\pm1\}^S} \#Y^{(q)}_{S,s}.$$

Here (r, c) *indexes the entry* $\mathbf{W}_{rc}$ *whose support changes at the transition,* $\sigma \in \{\pm 1\}$ *is its one-sided sign,* $S \subseteq \{1, \dots, d\}^2$ *is the active support away from* (r, c)*, and* $s \in \{\pm 1\}^S$ *is the corresponding active sign pattern; the sets* $Z^{(q)}_{(r,c),\sigma,S,s}$ *and* $Y^{(q)}_{S,s}$ *are defined in Appendix A.2.*

Topological Transitions. For exact NOTEARS and DAGMA, different topological orders compatible with the same signed directed support are alternative representations, not distinct branches. At a transition, we first remove zero-weight edges and recompute reachability. An inactive edge $i \to j$ is graph-feasible precisely when there is no directed path from j to i in the remaining graph. Deletion may expose previously forbidden directions, so their inactive stationarity inequalities must also be checked. Only graph-feasible inactive edges whose inequalities are tight can enter continuously. Tied entries are examined jointly: a successor must be acyclic and satisfy the one-sided active-sign and inactive-stationarity conditions. When multiple regular successors are admissible, the single-path implementation chooses the outgoing tangent closest to the incoming tangent, with a fixed lexicographic rule for exact ties. If no regular successor can be certified, the last one-sided solution initializes the warm-started corrector. These checks concern the directed support rather than an arbitrarily stored topological order; GOLEM, with a soft acyclicity penalty, has no hard topological transition.

7 Experiments and Results

We compare `DAG-flow` against the standard *sequential warm-start retraining* baseline, denoted Grid-WS-G: the underlying objective is reoptimized at G evenly spaced points in the same chosen λ-interval, each warm-started from the previous solution. All experiments cover three continuous DAG learning paradigms: NOTEARS, GOLEM (under λ_1-path and λ_2-path regimes, denoted as GOLEM1 and GOLEM2), and DAGMA.

Path Exactness and Scalability. We first test path-following accuracy and efficiency on three synthetic linear-SEM data-generating processes (DGPs). Given a stationary solution at an initial λ, both methods are driven toward $\lambda + 0.1\lambda_{\max}$ under a matched wall-clock budget. Table 1 reports stationarity residuals on DGP-1 at a fixed compute cutoff. It shows that `DAG-flow` lands exactly on the stationary path across all objectives and the full sweep of initial λ, whereas Grid-WS remains far from convergence under the same budget. The restart column shows that only 9-27% of transition points triggered a warm restart: the vast majority were resolved in closed form by the threshold conditions and an active-set refresh, with the small residual fraction reflecting unavoidable stationary-branch discontinuities. Figure 2 compares the wall-clock time for computing the entire path as either the sample size n or the number of nodes d varies. The advantage of `DAG-flow` becomes more pronounced as the problem size grows. This is consistent with the theory since one analytic segment represents a continuum of regularization values, whereas Grid-WS-G repeatedly solves nonconvex optimization problems at isolated grid points. Full implementation details, and additional results on two further synthetic data-generating processes are deferred to Appendices M and N.

Table 1. Active stationarity residuals of `DAG-flow` and Grid-WS-1 on DGP-1 at the finite-time cutoff (lower is better). For each row, the initial solution is obtained at the listed λ, and the target parameter is $\lambda + 0.1\,\lambda_{\max}$. The rightmost column reports the average percentage of TPs that triggered a warm-restart in the full path-following until $\lambda_{\max}$. GOLEM1 uses $\lambda_2 = 0.02$ and GOLEM2 uses $\lambda_1 = 0.005$. Values below 10^{-6} are reported as 0.

Initial λ	Seed	NOTEARS		GOLEM1		GOLEM2		DAGMA		Avg. Restart
		`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	
$0.10\lambda_{\max}$	42	0.00	0.721(0.108)	0.00	0.314(0.036)	0.00	0.382(0.049)	0.00	0.693(0.084)	18%
$0.15\lambda_{\max}$		0.00	0.648(0.071)	0.00	0.282(0.026)	0.00	0.341(0.064)	0.00	0.611(0.075)	22%
$0.20\lambda_{\max}$		0.00	0.573(0.092)	0.00	0.259(0.041)	0.00	0.318(0.033)	0.00	0.557(0.062)	19%
$0.25\lambda_{\max}$		0.00	0.526(0.052)	0.00	0.271(0.058)	0.00	0.287(0.021)	0.00	0.489(0.091)	25%
$0.30\lambda_{\max}$		0.00	0.471(0.083)	0.00	0.221(0.020)	0.00	0.296(0.057)	0.00	0.444(0.047)	21%
$0.35\lambda_{\max}$		0.00	0.515(0.132)	0.00	0.239(0.049)	0.00	0.263(0.028)	0.00	0.468(0.104)	27%
$0.40\lambda_{\max}$		0.00	0.389(0.038)	0.00	0.201(0.018)	0.00	0.231(0.036)	0.00	0.373(0.056)	20%
$0.45\lambda_{\max}$		0.00	0.342(0.067)	0.00	0.188(0.032)	0.00	0.246(0.048)	0.00	0.325(0.030)	17%
$0.50\lambda_{\max}$		0.00	0.297(0.026)	0.00	0.169(0.015)	0.00	0.209(0.024)	0.00	0.284(0.052)	15%
$0.55\lambda_{\max}$		0.00	0.254(0.051)	0.00	0.151(0.024)	0.00	0.191(0.041)	0.00	0.247(0.027)	14%
$0.60\lambda_{\max}$		0.00	0.232(0.023)	0.00	0.163(0.034)	0.00	0.174(0.018)	0.00	0.221(0.039)	13%
$0.65\lambda_{\max}$		0.00	0.188(0.039)	0.00	0.127(0.013)	0.00	0.158(0.027)	0.00	0.186(0.020)	11%
$0.70\lambda_{\max}$		0.00	0.169(0.019)	0.00	0.118(0.021)	0.00	0.146(0.016)	0.00	0.154(0.034)	9%
$0.10\lambda_{\max}$	2026	0.00	0.682(0.089)	0.00	0.337(0.045)	0.00	0.361(0.035)	0.00	0.729(0.126)	16%
$0.15\lambda_{\max}$		0.00	0.691(0.115)	0.00	0.301(0.031)	0.00	0.354(0.075)	0.00	0.602(0.049)	21%
$0.20\lambda_{\max}$		0.00	0.541(0.048)	0.00	0.276(0.056)	0.00	0.329(0.031)	0.00	0.512(0.070)	20%
$0.25\lambda_{\max}$		0.00	0.558(0.088)	0.00	0.238(0.023)	0.00	0.303(0.061)	0.00	0.471(0.045)	23%
$0.30\lambda_{\max}$		0.00	0.439(0.041)	0.00	0.247(0.046)	0.00	0.276(0.024)	0.00	0.459(0.073)	18%
$0.35\lambda_{\max}$		0.00	0.486(0.082)	0.00	0.217(0.027)	0.00	0.279(0.051)	0.00	0.431(0.055)	24%
$0.40\lambda_{\max}$		0.00	0.407(0.064)	0.00	0.184(0.017)	0.00	0.252(0.044)	0.00	0.361(0.035)	22%
$0.45\lambda_{\max}$		0.00	0.316(0.029)	0.00	0.207(0.040)	0.00	0.219(0.022)	0.00	0.348(0.066)	18%
$0.50\lambda_{\max}$		0.00	0.329(0.061)	0.00	0.158(0.017)	0.00	0.228(0.046)	0.00	0.268(0.031)	16%
$0.55\lambda_{\max}$		0.00	0.236(0.028)	0.00	0.159(0.030)	0.00	0.177(0.020)	0.00	0.259(0.049)	14%
$0.60\lambda_{\max}$		0.00	0.247(0.047)	0.00	0.143(0.015)	0.00	0.188(0.038)	0.00	0.210(0.023)	12%
$0.65\lambda_{\max}$		0.00	0.174(0.021)	0.00	0.136(0.026)	0.00	0.148(0.016)	0.00	0.192(0.041)	10%
$0.70\lambda_{\max}$		0.00	0.181(0.035)	0.00	0.109(0.012)	0.00	0.162(0.029)	0.00	0.140(0.018)	9%

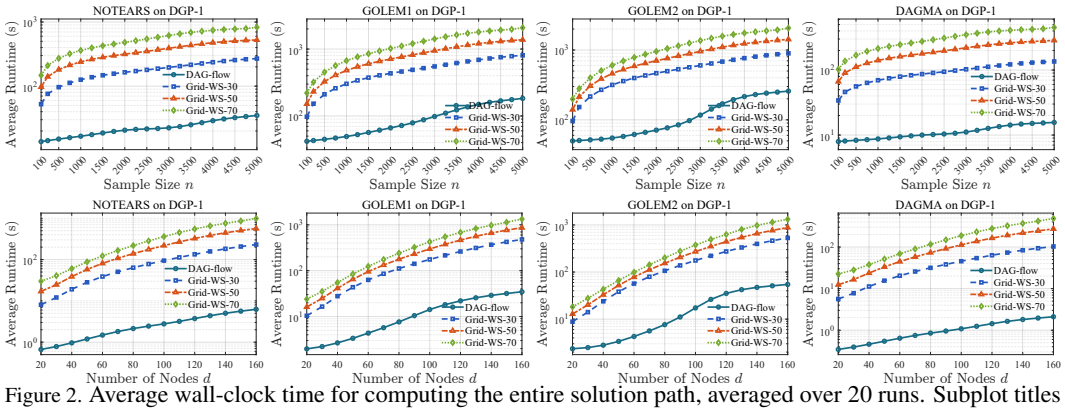


Figure 2. Average wall-clock time for computing the entire solution path, averaged over 20 runs. Subplot titles indicate the experimental settings; when one of n and d is varied, the other dimension is held fixed.

Table 2. Results of Test NLL and average runtime (Time) of `DAG-flow` and Grid-WS-50 on real or semi-real public benchmarks. Lower NLL indicates better out-of-sample fit. Results are reported as mean(std) over 20 random 70/20/10 train/validation/test splits for Test NLL, while Time is the computational runtime in seconds.

Dataset	NOTEARS				GOLEM				DAGMA			
	`DAG-flow`		Grid-WS		`DAG-flow`		Grid-WS		`DAG-flow`		Grid-WS	
	Test NLL	Time	Test NLL	Time	Test NLL	Time	Test NLL	Time	Test NLL	Time	Test NLL	Time
Sachs	0.986(0.037)	42.8	1.01(0.043)	146	0.958(0.034)	168	0.998(0.041)	359	0.991(0.038)	47.6	1.02(0.044)	171
DREAM4-10	1.22(0.092)	9.84	1.29(0.108)	71.5	1.16(0.085)	46.2	1.27(0.104)	188	1.20(0.090)	12.7	1.25(0.101)	78.3
DREAM4-100MF	1.78(0.165)	845	1.92(0.193)	7.42×10^3	1.69(0.154)	3.91×10^3	1.85(0.181)	1.48×10^4	1.73(0.160)	1.13×10^3	1.89(0.187)	9.35×10^3
SynTReN-300	1.43(0.057)	1.52×10^4	1.47(0.066)	9.85×10^4	1.36(0.052)	6.12×10^4	1.42(0.063)	1.84×10^5	1.40(0.055)	1.90×10^4	1.46(0.065)	1.20×10^5

Public Benchmarks. We further evaluate on the real-world flow-cytometry dataset Sachs [Sachs et al., 2005] together with three in-silico gene-regulatory benchmarks: DREAM4-10 and DREAM4-100MF [Schaffter et al., 2011, Guo et al., 2016], and SynTReN-300 [Van den Bulcke et al., 2006]. Following standard practice in likelihood-based DAG learning when reliable ground-truth graphs are unavailable [Brouillard et al., 2020, Li et al., 2024], we score each candidate graph by refitting an unregularized linear SEM on its support, selecting via validation Negative Log-Likelihood (NLL), and reporting test NLL on a held-out split. As shown in Table 2, `DAG-flow` achieves lower Test NLL than Grid-WS-50 across all dataset/model pairs while reducing runtime by roughly 3-10×. The

advantage is largest on the high-dimensional benchmarks DREAM4-100MF and SynTReN-300, where Grid-WS-50 pays a full nonconvex solve multiple times. The accuracy gain stems from `DAG-flow` exposing a structured continuum of candidate graphs from which validation can select, rather than the coarse, predetermined sample produced by a fixed grid.

8 Discussion

In this work, we propose `DAG-flow`, an exact regularization path-following framework for continuous DAG learning. Empirically, `DAG-flow` outperforms grid-search baselines at a fraction of the compute and consistently improves structural stability. We view this work as an algorithmic interface that opens the door to downstream applications such as model selection and ensembling.

Limitations & Future Directions. Our path-following framework `DAG-flow` tracks stationary branches initialized from a given solution, but does not by itself guarantee coverage of all stationary branches in highly non-convex landscapes. Discontinuous transitions and turning points may still require warm-started reoptimization, although our experiments show that such events are relatively infrequent. In addition, our theory mainly focuses on linear SEMs and one-dimensional regularization paths; extending the analysis to nonlinear DAG models, nonparametric objectives, and genuinely multi-parameter continuation remains an important direction.

Acknowledgments

This work was supported in part by NSF grant 2212175 and NIH grants RF1AG072449, RF1AG084178, R01AG080991, R01AG080624, R01AG076448, and R01AG076234. This work was also partially supported by an Amazon Web Services compute grant and the NVIDIA Academic Grant Program.

References

Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse gaussian bayesian networks. *The Journal of Machine Learning Research*, 16(1):2273–2328, 2015.

Bryon Aragam, Jiaying Gu, and Qing Zhou. Learning large-scale bayesian networks with the sparsebn package. *Journal of Statistical Software*, 91:1–38, 2019.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 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.

Richard A Brualdi, Herbert John Ryser, et al. *Combinatorial matrix theory*, volume 39. Springer, 1991.

John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330, 2004.

Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Global optimality in bivariate gradient-based dag learning. *Advances in Neural Information Processing Systems*, 36:17929–17968, 2023.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. 2004.

Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, Junyang Qian, and James Yang. glmnet: Lasso and elastic-net regularized generalized linear models. *Astrophysics Source Code Library*, page ascl:2308.011, 2023. URL https://ascl.net/2308.011.

Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22, 2010.

Fei Fu and Qing Zhou. Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501): 288–300, 2013.

Jiaying Gu, Fei Fu, and Qing Zhou. Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, 29(1):161–176, 2019.

Shun Guo, Qingshan Jiang, Lifei Chen, and Donghui Guo. Gene regulatory network inference using pls-based methods. *BMC bioinformatics*, 17(1):545, 2016.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*, volume 1. Springer, 2016.

Utkarsh Mahadeo Khaire and R Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1060–1073, 2022.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Hangjian Li, Oscar Hernan Madrid Padilla, and Qing Zhou. Learning gaussian dags from network data. *Journal of Machine Learning Research*, 25(377):1–52, 2024.

Xiang Li, Shanghong Xie, Peter McColgan, Sarah J Tabrizi, Rachael I Scahill, Donglin Zeng, and Yuanjia Wang. Learning subject-specific directed acyclic graphs with mixed effects structural equation models from observational data. *Frontiers in genetics*, 9:430, 2018.

Sanjay Manjunath, Yujia Zheng, Samuel Horváth, and Kun Zhang. Top-k structure search with solution path. 2025. URL https://openreview.net/forum?id=FNA0AlYBMS.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2014.

Marcello Massimo Negri, Fabricio Arend Torres, and Volker Roth. Conditional matrix flows for gaussian graphical models. *Advances in Neural Information Processing Systems*, 36:25095–25111, 2023.

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.

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8176–8198. Pmlr, 2022.

Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):659–677, 2007.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

Seyed Saman Saboksayr, Gonzalo Mateos, and Mariano Tepper. ColiDE: Concomitant linear DAG estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fGAIgO75dG.

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.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011.

Mark Schmidt, Alexandru Niculescu-Mizil, Kevin Murphy, et al. Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pages 1278–1283, 2007.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of statistics*, 39(3):1335–1371, 2011.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7 (1):43, 2006.

Diego Vidaurre, Concha Bielza, and Pedro Larrañaga. Learning an l1-regularized gaussian bayesian network in the equivalence class space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5):1231–1242, 2010.

Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. *Advances in Neural Information Processing Systems*, 33:3895–3906, 2020.

Danru Xu, Erdun Gao, Wei Huang, Menghan Wang, Andy Song, and Mingming Gong. On the sparse dag structure learning based on adaptive lasso. *arXiv preprint arXiv:2209.02946*, 2022.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Mingming Gong, Biwei Huang, Kun Zhang, Anton van den Hengel, and Javen Qinfeng Shi. Analytic dag constraints for differentiable dag learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International conference on artificial intelligence and statistics*, pages 3414–3425. Pmlr, 2020.

Hua Zhou and Kenneth Lange. Path following in the exact penalty method of convex programming. *Computational optimization and applications*, 61(3):609–634, 2015.

Contents

1 Introduction **1**
- 1.1 Related Work . . . 3

2 Preliminaries **3**

3 Regularization Path for NOTEARS **4**

4 Regularization Path for GOLEM **5**

5 Regularization Path for DAGMA **7**

6 Algorithmic Overview **7**

7 Experiments and Results **8**

8 Discussion **10**

A Missing Definitions **14**
- A.1 Terms in Theorem 4.4 . . . 14
- A.2 Terms in Theorem 6.1 . . . 15

B Pseudocode **16**

C Proof of Theorem 3.1 **17**

D Proof of Theorem 3.2 **21**

E Proof of Theorem 3.3 **28**

F Proof of Theorem 3.4 **31**

G Proof of Theorem 4.1 **35**

H Proof of Theorem 4.2 **41**

I Proof of Theorem 4.3 **48**

J Proof of Theorem 4.4 **57**

K Proof of Theorem 5.1 **57**

L Proof of Theorem 6.1 **61**
- L.1 Part I . . . 61
- L.2 Part II . . . 67

M Experimental Details **72**
- M.1 Dataset Details . . . 72
- M.2 Implementation . . . 73

N Additional Empirical Results **74**

A Missing Definitions

Appendix A provides the detailed definitions deferred from the main text due to space constraints.

A.1 Terms in Theorem 4.4

Throughout, $\mathcal{A}$ denotes the active support along the current segment, $\mathcal{A}^c$ its complement.

λ_1-path. With $\lambda_2 = \lambda_2^0$ held fixed, the four candidates are defined as follows.

$$\lambda_{1,\mathrm{drop}} := \inf\left\{\lambda_1 \in (\lambda_1^0, \lambda_{\max}) : \exists j \in \{1,\dots,d\},\ \exists i \in \mathcal{A}_j \text{ s.t. } \mathbf{W}_{ij}(\lambda_1) = 0\right\},$$

$$\lambda_{1,\mathrm{enter}} := \inf\left\{\lambda_1 \in (\lambda_1^0, \lambda_{\max}) : \exists j \in \{1,\dots,d\},\ \exists i \notin \mathcal{A}_j \text{ s.t. } \left|\frac{d\,[\mathbf{X}^\top\mathbf{X}(\mathbf{W}-\mathbf{I})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-\top}]_{ij}\right| = \lambda_1\right\},$$

$$\lambda_{1,\mathrm{dom}} := \inf\left\{\lambda_1 \in (\lambda_1^0, \lambda_{\max}) : \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 = 0 \text{ or } \det(\mathbf{I}-\mathbf{W}) = 0\right\},$$

$$\lambda_{1,\mathrm{turn}} := \inf\left\{\lambda_1 \in (\lambda_1^0, \lambda_{\max}) : \exists \mathbf{U} \neq \mathbf{0},\ \mathbf{U}_{\mathcal{A}^c} = \mathbf{0},\ \text{satisfying (T1)}\right\}.$$

The turning-point candidate $\lambda_{1,\mathrm{turn}}$ is the smallest λ_1 at which the linear operator governing the velocity acquires a nontrivial kernel, and $\mathbf{U}$ satisfy

$$\begin{aligned}
&\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}\mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4}\sum_{r=1}^{d}\left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r}-\mathbf{X}_{:,r})\right]^\top \mathbf{U}_{\mathcal{A}_r,r} \\
&+ \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{U}^\top(\mathbf{I}-\mathbf{W})^{-\top}\right]_{\mathcal{A}_j,j} \\
&+ 2\lambda_2^0\left\{\mathbf{U}_{\mathcal{A}_j,j} \odot \left[(e^{\mathbf{W}\odot\mathbf{W}})^\top\right]_{\mathcal{A}_j,j} + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-t)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{t(\mathbf{W}\odot\mathbf{W})}\,dt\right)^\top\right]_{\mathcal{A}_j,j}\right\} = \mathbf{0},\ \ j=1,\dots,d.
\end{aligned} \tag{T1}$$

λ_2-path. With $\lambda_1 = \lambda_1^0$ held fixed, the candidates are

$$\lambda_{2,\mathrm{drop}} := \inf\left\{\lambda_2 \in (\lambda_2^0, +\infty) : \exists j \in \{1,\dots,d\},\ \exists i \in \mathcal{A}_j \text{ s.t. } \mathbf{W}_{ij}(\lambda_2) = 0\right\},$$

$$\lambda_{2,\mathrm{enter}} := \inf\left\{\lambda_2 \in (\lambda_2^0, +\infty) : \exists j \in \{1,\dots,d\},\ \exists i \notin \mathcal{A}_j \text{ s.t. } \left|\frac{d\,[\mathbf{X}^\top\mathbf{X}(\mathbf{W}-\mathbf{I})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-\top}]_{ij}\right| = \lambda_1^0\right\},$$

$$\lambda_{2,\mathrm{dom}} := \inf\left\{\lambda_2 \in (\lambda_2^0, +\infty) : \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 = 0 \text{ or } \det(\mathbf{I}-\mathbf{W}) = 0\right\},$$

$$\lambda_{2,\mathrm{turn}} := \inf\left\{\lambda_2 \in (\lambda_2^0, +\infty) : \exists \mathbf{U} \neq \mathbf{0},\ \mathbf{U}_{\mathcal{A}^c} = \mathbf{0},\ \text{satisfying (T2)}\right\},$$

where (T2) is the corresponding homogeneous velocity system at the running λ_2,

$$\begin{aligned}
&\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}\mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4}\sum_{r=1}^{d}\left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r}-\mathbf{X}_{:,r})\right]^\top \mathbf{U}_{\mathcal{A}_r,r} \\
&+ \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{U}^\top(\mathbf{I}-\mathbf{W})^{-\top}\right]_{\mathcal{A}_j,j} \\
&+ 2\lambda_2\left\{\mathbf{U}_{\mathcal{A}_j,j} \odot \left[(e^{\mathbf{W}\odot\mathbf{W}})^\top\right]_{\mathcal{A}_j,j} + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-t)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{t(\mathbf{W}\odot\mathbf{W})}\,dt\right)^\top\right]_{\mathcal{A}_j,j}\right\} = \mathbf{0},\ \ j=1,\dots,d.
\end{aligned} \tag{T2}$$

A.2 Terms in Theorem 6.1

λ_1-path: support / sign wall transitions.

$$Z^{(1)}_{(r,c),\sigma,S,s} := \left\{ (\mathbf{W},\lambda_1) : \begin{cases} \lambda_1 = -\sigma\left[\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{rc}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{rc}\right], \qquad 0<\lambda_1<\lambda_{\max}, \\ \|\mathbf{W}\|_F \le M, \qquad \mathbf{W}_{rc}=0, \qquad \mathbf{W}_{ij}=0, \quad \forall (i,j)\notin S\cup\{(r,c)\}, \\ \dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2^0\mathbf{W}_{ij}\left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji} \\ \qquad\qquad - \sigma s_{ij}\left[\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{rc}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{rc}\right] = 0, \quad \forall (i,j)\in S, \\ s_{ij}\mathbf{W}_{ij} > 0, \quad \forall (i,j)\in S, \\ \left|\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}\right| \le -\sigma\left[\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{rc}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{rc}\right], \\ \qquad\qquad \forall (i,j)\notin S\cup\{(r,c)\}. \end{cases} \right\}.$$

λ_1-path: singular regime within a fixed support.

$$Y^{(1)}_{S,s} := \left\{ (\mathbf{W},\lambda_1) : \begin{cases} 0<\lambda_1<\lambda_{\max}, \qquad \|\mathbf{W}\|_F \le M, \\ \mathbf{W}_{ij}=0, \quad \forall (i,j)\notin S, \\ \dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2^0\mathbf{W}_{ij}\left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji} + \lambda_1 s_{ij} = 0, \quad \forall (i,j)\in S, \\ s_{ij}\mathbf{W}_{ij} > 0, \quad \forall (i,j)\in S, \\ \left|\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}\right| < \lambda_1, \quad \forall (i,j)\notin S, \\ \det\left(\left[\dfrac{\partial}{\partial \mathbf{W}_{uv}}\left\{\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2^0\mathbf{W}_{ij}\left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji}\right\}\right]_{(i,j),(u,v)\in S}\right) = 0. \end{cases} \right\}.$$

λ_2-path: support / sign wall transitions.

$$Z^{(2)}_{(r,c),\sigma,S,s} := \left\{ (\mathbf{W},\lambda_2) : \begin{cases} 0<\lambda_2<\lambda_2^{\max}, \qquad \|\mathbf{W}\|_F \le M, \\ \mathbf{W}_{rc}=0, \qquad \mathbf{W}_{ij}=0, \quad \forall (i,j)\notin S\cup\{(r,c)\}, \\ \dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{rc}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{rc} + \lambda_1^0\sigma = 0, \\ 2\lambda_2\mathbf{W}_{ij}\left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji} = -\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} - \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} - \lambda_1^0 s_{ij}, \\ \qquad\qquad \forall (i,j)\in S, \\ s_{ij}\mathbf{W}_{ij} > 0, \quad \forall (i,j)\in S, \\ \left|\dfrac{d\left[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\right]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}\right| \le \lambda_1^0, \\ \qquad\qquad \forall (i,j)\notin S\cup\{(r,c)\}. \end{cases} \right\}.$$

λ_2-path: singular regime within a fixed support.

$$Y^{(2)}_{S,s} := \left\{ (\mathbf{W}, \lambda_2) : \begin{cases} 0 < \lambda_2 < \lambda_2^{\max}, \qquad \|\mathbf{W}\|_F \le M, \\ \mathbf{W}_{ij} = 0, \quad \forall (i,j) \notin S, \\ \dfrac{d\left[\mathbf{X}^\top(\mathbf{X}\mathbf{W} - \mathbf{X})\right]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + \left[(\mathbf{I} - \mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2 \mathbf{W}_{ij} \left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji} + \lambda_1^0 s_{ij} = 0, \quad \forall (i,j) \in S, \\ s_{ij}\mathbf{W}_{ij} > 0, \quad \forall (i,j) \in S, \\ \left| \dfrac{d\left[\mathbf{X}^\top(\mathbf{X}\mathbf{W} - \mathbf{X})\right]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + \left[(\mathbf{I} - \mathbf{W})^{-\top}\right]_{ij} \right| < \lambda_1^0, \quad \forall (i,j) \notin S, \\ \det\left(\left[\dfrac{\partial}{\partial \mathbf{W}_{uv}} \left\{ \dfrac{d\left[\mathbf{X}^\top(\mathbf{X}\mathbf{W} - \mathbf{X})\right]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + \left[(\mathbf{I} - \mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2 \mathbf{W}_{ij} \left(e^{\mathbf{W}\odot\mathbf{W}}\right)_{ji} \right\} \right]_{(i,j),(u,v)\in S} \right) = 0. \end{cases} \right\}.$$

The determinant is computed with respect to any fixed ordering of the entries of S; the value of the determinant is invariant up to sign under reorderings, so the zero set itself is well-defined.

B Pseudocode

We give the full pseudocode of the regularization path algorithm from Section 6, separated into the piecewise-linear case (Algorithm 1) and the piecewise-smooth case (Algorithm 2).

Algorithm 1: `DAG-flow` for piecewise-affine paths

Input : Data matrix $\mathbf{X}$; objective $\mathcal{M} \in \{(1), (4)\}$; initial stationary point $\mathbf{W}(\lambda_0)$ at λ_0.
Output : Regularization path $\{\mathbf{W}(\lambda) : \lambda \in (0, \lambda_{\max}]\}$.

Compute $\lambda_{\max}$ by Theorem 3.1 if $\mathcal{M} = (1)$, and by Theorem 5.1 if $\mathcal{M} = (4)$.
for $\lambda : \lambda_0 \to \lambda_{\max}$ *and* $\lambda : \lambda_0 \to 0$ **do**
  $\lambda \leftarrow \lambda_0, \quad \mathbf{W} \leftarrow \mathbf{W}(\lambda_0)$.
  while $\lambda \in (0, \lambda_{\max})$ **do**
    $\mathcal{A} \leftarrow \{(i,j) : \mathbf{W}_{ij} \neq 0\}, \quad \mathcal{A}_j \leftarrow \{i : (i,j) \in \mathcal{A}\}$.
    Freeze $\operatorname{sign}(\mathbf{W}_{\mathcal{A}})$ and the current acyclic stratum.
    Solve the system in Theorem 3.3 for NOTEARS, or in Corollary 5.3 for DAGMA, to obtain $\dot{\mathbf{W}} = d\mathbf{W}/d\lambda$, with $\dot{\mathbf{W}}_{\mathcal{A}^c} \leftarrow \mathbf{0}$.
    Solve λ_{next} in the current direction by Theorem 3.4, including endpoint 0 or $\lambda_{\max}$.
    Record $\mathbf{W}(t) = \mathbf{W} + (t - \lambda)\dot{\mathbf{W}}$, for $t \in [\min\{\lambda, \lambda_{\text{next}}\}, \max\{\lambda, \lambda_{\text{next}}\}]$.
    $\mathbf{W} \leftarrow \mathbf{W} + (\lambda_{\text{next}} - \lambda)\dot{\mathbf{W}}, \quad \lambda \leftarrow \lambda_{\text{next}}$.
    if $\lambda \in \{0, \lambda_{\max}\}$ **then**
      break
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{(i,j) : \mathbf{W}_{ij} = 0\} \cup \{(i,j) : \lambda_{ij}^{\text{enter}} = \lambda\}$.
    Assign one-sided signs to newly inserted entries and update the acyclic stratum.
    if $\mathbf{W}$ *is not stationary at current* λ **then**
      Warm-start at current λ from $\mathbf{W}$.
return $\{\mathbf{W}(\lambda) : \lambda \in (0, \lambda_{\max}]\}$.

Algorithm 2: `DAG-flow` for piecewise-smooth GOLEM paths

Input : Data matrix $\mathbf{X}$; initial stationary point $\mathbf{W}(\lambda_1^0, \lambda_2^0)$; path type $q \in \{1, 2\}$. For the λ_2-path, also provide the endpoint $\lambda_2^{\max}$.
Output : Regularization path of the selected GOLEM (3) branch.

if $q = 1$ **then**
  Fix $\lambda_2 = \lambda_2^0$, set $\lambda_q = \lambda_1$, and compute the endpoint by Theorem 4.1.
else
  Fix $\lambda_1 = \lambda_1^0$, set $\lambda_q = \lambda_2$, and use endpoint $\lambda_2^{\max}$.
for $\lambda_q : \lambda_q^0 \to$ *upper endpoint and* $\lambda_q : \lambda_q^0 \to 0$ **do**
  $\lambda_q \leftarrow \lambda_q^0, \quad \mathbf{W} \leftarrow \mathbf{W}(\lambda_1^0, \lambda_2^0)$.
  while λ_q *lies in the selected interval* **do**
    $\mathcal{A} \leftarrow \{(i,j) : \mathbf{W}_{ij} \neq 0\}, \quad \mathcal{A}_j \leftarrow \{i : (i,j) \in \mathcal{A}\}$.
    Freeze $\operatorname{sign}(\mathbf{W}_{\mathcal{A}})$.
    Solve the linear system in Theorem 4.3 for $\mathbf{V}^{(q)}$, with $\frac{d\mathbf{W}_{\mathcal{A}_j,j}}{d\lambda_q} = \mathbf{V}^{(q)}_{\mathcal{A}_j,j}$ and $\frac{d\mathbf{W}_{\mathcal{A}^c}}{d\lambda_q} = \mathbf{0}$.
    Numerically integrate the ODE until the first event in Theorem 4.4 or the selected endpoint.
    Record the segment of $\mathbf{W}(\lambda_1, \lambda_2^0)$ if $q = 1$, or $\mathbf{W}(\lambda_1^0, \lambda_2)$ if $q = 2$.
    $\lambda_q \leftarrow \lambda_{q,\text{next}}, \quad \mathbf{W} \leftarrow \mathbf{W}(\lambda_{q,\text{next}})$.
    if $\lambda_q >$ *selected endpoint* **then**
      break
    if $\lambda_{q,\text{next}} = \lambda_{q,\text{dom}}$ **then**
      break
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{(i,j) : \mathbf{W}_{ij} = 0\} \cup \{(i,j) : \lambda_{q,ij}^{\text{enter}} = \lambda_q\}$.
    Assign one-sided signs to newly inserted entries.
    if $\lambda_{q,\text{next}} = \lambda_{q,\text{turn}}$ *or stationarity fails* **then**
      Warm-start (3) at current (λ_1, λ_2) from $\mathbf{W}$.
return $\{\mathbf{W}(\lambda_1, \lambda_2^0)\}$ *if* $q = 1$, *or* $\{\mathbf{W}(\lambda_1^0, \lambda_2)\}$ *if* $q = 2$.

C Proof of Theorem 3.1

Proof. Let $\mathcal{P}_d$ denote the set of all $d \times d$ permutation matrices. We first record the precise feasible-set structure induced by the NOTEARS constraint. For every $\mathbf{W} \in \mathbb{R}^{d\times d}$, we know

$$\begin{aligned} h(\mathbf{W}) &= \operatorname{tr}\left[\mathbf{I} + (\mathbf{W} \odot \mathbf{W}) + \frac{1}{2!}(\mathbf{W} \odot \mathbf{W})^2 + \frac{1}{3!}(\mathbf{W} \odot \mathbf{W})^3 + \cdots\right] - d \\ &= \sum_{i=1}^{d} W_{ii}^2 + \sum_{k=2}^{\infty} \frac{1}{k!} \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \cdots \sum_{i_k=1}^{d} W_{i_1 i_2}^2 W_{i_2 i_3}^2 \cdots W_{i_{k-1} i_k}^2 W_{i_k i_1}^2 \geq 0. \end{aligned} \tag{5}$$

Thus, if the directed support of $\mathbf{W}$ contained a directed cycle $i_1 \to i_2 \to \cdots \to i_r \to i_1$, then the strictly positive monomial $W_{i_1 i_2}^2 W_{i_2 i_3}^2 \cdots W_{i_{r-1} i_r}^2 W_{i_r i_1}^2$ would appear in (5) and hence $h(\mathbf{W}) > 0$. Therefore, on $\{h = 0\}$ the directed support of any feasible $\mathbf{W}$ admits a topological ordering [Brualdi et al., 1991]; equivalently, there exists $\mathbf{P} \in \mathcal{P}_d$ such that $(\mathbf{P}^\top \mathbf{W} \mathbf{P})_{ab} = 0$ whenever $1 \leq b \leq a \leq d$. Conversely, if for some $\mathbf{P} \in \mathcal{P}_d$ the matrix $\mathbf{P}^\top \mathbf{W} \mathbf{P}$ is strictly upper triangular, then $\mathbf{P}^\top (\mathbf{W} \odot \mathbf{W}) \mathbf{P} = (\mathbf{P}^\top \mathbf{W} \mathbf{P}) \odot (\mathbf{P}^\top \mathbf{W} \mathbf{P})$ is strictly upper triangular and therefore nilpotent. Hence

$$e^{\mathbf{P}^\top (\mathbf{W}\odot\mathbf{W})\mathbf{P}} = \mathbf{I} + \mathbf{P}^\top (\mathbf{W} \odot \mathbf{W}) \mathbf{P} + \frac{1}{2!}\left[\mathbf{P}^\top (\mathbf{W} \odot \mathbf{W}) \mathbf{P}\right]^2 + \cdots$$

is upper triangular with every diagonal entry equal to one. By similarity invariance of the exponential and the trace, $\operatorname{tr}\left(e^{\mathbf{W}\odot\mathbf{W}}\right) = \operatorname{tr}\left(e^{\mathbf{P}^\top (\mathbf{W}\odot\mathbf{W})\mathbf{P}}\right) = d$, and hence

$$\left\{\mathbf{W} \in \mathbb{R}^{d\times d} : h(\mathbf{W}) = 0\right\} = \bigcup_{\mathbf{P}\in\mathcal{P}_d} \left\{\mathbf{W} \in \mathbb{R}^{d\times d} : (\mathbf{P}^\top \mathbf{W} \mathbf{P})_{ab} = 0 \text{ whenever } 1 \leq b \leq a \leq d\right\}.$$

Now fix any feasible $\mathbf{W}$, and choose $\mathbf{P} \in \mathcal{P}_d$ such that $\mathbf{P}^\top \mathbf{W}\mathbf{P}$ is strictly upper triangular. For every $t \in [0,1]$, the entire ray $t\mathbf{W}$ remains feasible because $\mathbf{P}^\top(t\mathbf{W})\mathbf{P} = t\,\mathbf{P}^\top \mathbf{W}\mathbf{P}$ is still strictly upper triangular. Along this ray, we have

$$
\begin{aligned}
&\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}(t\mathbf{W})\|_F^2 + \lambda\|t\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \\
&= \frac{1}{2n}\operatorname{tr}\big[(\mathbf{X} - t\mathbf{X}\mathbf{W})^\top(\mathbf{X} - t\mathbf{X}\mathbf{W}) - \mathbf{X}^\top\mathbf{X}\big] + \lambda t\|\mathbf{W}\|_1 \\
&= -\frac{t}{2}\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) - \frac{t}{2}\operatorname{tr}\left(\mathbf{W}^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right) + \frac{t^2}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 + \lambda t\|\mathbf{W}\|_1 \\
&= -t\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) + \frac{t^2}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 + \lambda t\|\mathbf{W}\|_1 \\
&= -t\operatorname{tr}\left[\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\,\mathbf{P}^\top\mathbf{W}\mathbf{P}\right] + \frac{t^2}{2n}\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F^2 + \lambda t\sum_{1\le a<b\le d}|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| \\
&= \frac{t^2}{2n}\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F^2 + t\sum_{1\le a<b\le d}\left[\lambda|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| - \left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}\right],
\end{aligned}
$$

where in the final equality we used that $\mathbf{P}^\top\mathbf{W}\mathbf{P}$ is strictly upper triangular and that $\mathbf{P}^\top(\mathbf{X}^\top\mathbf{X}/n)\mathbf{P}$ is symmetric, so the coefficient of $(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}$ in the trace is $(\mathbf{P}^\top(\mathbf{X}^\top\mathbf{X}/n)\mathbf{P})_{ab}$. The endpoint $t = 1$ gives the exact raywise chamber identity

$$
\begin{aligned}
&\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \\
&= \frac{1}{2n}\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F^2 + \sum_{1\le a<b\le d}\left[\lambda|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| - \left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}\right].
\end{aligned}
\tag{6}
$$

Since every strictly upper-triangular entry of $\mathbf{P}^\top(\mathbf{X}^\top\mathbf{X}/n)\mathbf{P}$ is an off-diagonal entry of $\mathbf{X}^\top\mathbf{X}/n$,

$$
\left|\left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}\right| \le \left\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \operatorname{diag}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\right\|_{\max} = \lambda_{\max} \quad (1 \le a < b \le d).
$$

Therefore the preceding identity (6) decomposes as

$$
\begin{aligned}
&\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \\
&= \frac{1}{2n}\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F^2 + (\lambda - \lambda_{\max})\sum_{1\le a<b\le d}|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| \\
&\quad + \sum_{1\le a<b\le d}\left[\lambda_{\max} - \left|\left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}\right|\right]|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| \\
&\quad + \sum_{1\le a<b\le d}\left[\left|\left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}\right||(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| - \left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}\right].
\end{aligned}
$$

Every term on the right-hand side is nonnegative whenever $\lambda \ge \lambda_{\max}$. Moreover,

$$
\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F = \|\mathbf{X}\mathbf{W}\|_F, \qquad \sum_{1\le a<b\le d}|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| = \|\mathbf{W}\|_1,
$$

because $\mathbf{P}^\top\mathbf{W}\mathbf{P}$ is strictly upper triangular. Hence, for every feasible $\mathbf{W}$,

$$
\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \ge \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 + (\lambda - \lambda_{\max})\|\mathbf{W}\|_1 \ge 0.
$$

Since $\mathbf{0}$ is feasible, $h(\mathbf{0}) = \operatorname{tr}(e^{\mathbf{0}}) - d = 0$, this proves that $\mathbf{0}$ is a global minimizer for every $\lambda \ge \lambda_{\max}$.

We next prove uniqueness. If $\lambda > \lambda_{\max}$ and $\mathbf{W} \neq \mathbf{0}$ is feasible, then $\|\mathbf{W}\|_1 > 0$, and the preceding bound yields

$$\begin{aligned}
&\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \\
&\geq \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 + (\lambda - \lambda_{\max})\|\mathbf{W}\|_1 \\
&> 0.
\end{aligned}$$

Thus no nonzero feasible point can attain the value at $\mathbf{0}$, so $\mathbf{0}$ is the unique global minimizer whenever $\lambda > \lambda_{\max}$.

Now suppose $\lambda = \lambda_{\max} > 0$, and let $\mathbf{W}$ be a feasible global minimizer. Choose $\mathbf{P} \in \mathcal{P}_d$ such that $\mathbf{P}^\top \mathbf{W}\mathbf{P}$ is strictly upper triangular. Since $\mathbf{0}$ is also feasible and globally minimizing, the exact nonnegative decomposition (6) gives

$$\begin{aligned}
0 &= \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_{\max}\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda_{\max}\|\mathbf{0}\|_1\right] \\
&= \frac{1}{2n}\left\|\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P})\right\|_F^2 + \sum_{1\leq a<b\leq d}\left[\lambda_{\max} - \left|\left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}\right|\right]|(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| \\
&\quad + \sum_{1\leq a<b\leq d}\left[\left|\left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}\right||(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}| - \left(\mathbf{P}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{P}\right)_{ab}(\mathbf{P}^\top\mathbf{W}\mathbf{P})_{ab}\right],
\end{aligned}$$

where each summand is nonnegative, so the first term must vanish, i.e., $\mathbf{X}\mathbf{P}(\mathbf{P}^\top\mathbf{W}\mathbf{P}) = \mathbf{0}$; equivalently, $\mathbf{X}\mathbf{W} = \mathbf{0}$. Returning to the exact objective-gap identity,

$$\begin{aligned}
0 &= \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_{\max}\|\mathbf{W}\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda_{\max}\|\mathbf{0}\|_1\right] \\
&= \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 - \operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) + \lambda_{\max}\|\mathbf{W}\|_1,
\end{aligned} \tag{7}$$

and using $\mathbf{X}\mathbf{W} = \mathbf{0}$, we obtain

$$\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) = \frac{1}{n}\operatorname{tr}\big(\mathbf{X}^\top(\mathbf{X}\mathbf{W})\big) = 0.$$

Thus $0 = \lambda_{\max}\|\mathbf{W}\|_1$. Since $\lambda_{\max} > 0$, it follows that $\|\mathbf{W}\|_1 = 0$, and therefore $\mathbf{W} = \mathbf{0}$. Hence $\mathbf{0}$ is also the unique global minimizer at the boundary $\lambda = \lambda_{\max} > 0$.

The same estimates yield a global convergence statement in the theorem. If $\lambda > \lambda_{\max}$ and $\{\mathbf{W}_m\}_{m\geq 1}$ is any feasible sequence satisfying

$$\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}_m\|_F^2 + \lambda\|\mathbf{W}_m\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right] \longrightarrow 0,$$

then we know

$$0 \leq \|\mathbf{W}_m\|_1 \leq \frac{\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}_m\|_F^2 + \lambda\|\mathbf{W}_m\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1\right]}{\lambda - \lambda_{\max}} \longrightarrow 0.$$

At the boundary $\lambda = \lambda_{\max} > 0$, if $\{\mathbf{W}_m\}_{m\geq 1}$ is feasible and the same objective gap with $\lambda_{\max}$ in place of λ tends to zero, then the nonnegative raywise decomposition gives

$$0 \leq \frac{1}{2n}\|\mathbf{X}\mathbf{W}_m\|_F^2 \leq \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}_m\|_F^2 + \lambda_{\max}\|\mathbf{W}_m\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda_{\max}\|\mathbf{0}\|_1\right],$$

and hence

$$\|\mathbf{X}\mathbf{W}_m\|_F \leq \sqrt{2n\left(\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}_m\|_F^2 + \lambda_{\max}\|\mathbf{W}_m\|_1\right] - \left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda_{\max}\|\mathbf{0}\|_1\right]\right)}.$$

Using the exact expansion of the objective gap (7),

$$
\begin{aligned}
\lambda_{\max}\|\mathbf{W}_m\|_1 &= \left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right] \\
&\quad -\frac{1}{2n}\|\mathbf{X}\mathbf{W}_m\|_F^2+\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_m\right) \\
&\le \left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right] \\
&\quad +\left|\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_m\right)\right| \\
&= \left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right] \\
&\quad +\frac{1}{n}\left|\operatorname{tr}\big(\mathbf{X}^\top(\mathbf{X}\mathbf{W}_m)\big)\right| \\
&\le \left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right] \\
&\quad +\frac{1}{n}\|\mathbf{X}\|_F\|\mathbf{X}\mathbf{W}_m\|_F \\
&\le \left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right] \\
&\quad +\frac{\|\mathbf{X}\|_F}{n}\sqrt{2n\left(\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_m\|_F^2+\lambda_{\max}\|\mathbf{W}_m\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda_{\max}\|\mathbf{0}\|_1\right]\right)}.
\end{aligned}
\tag{8}
$$

Dividing by $\lambda_{\max} > 0$ shows $\|\mathbf{W}_m\|_1 \to 0$. This strengthens the uniqueness conclusion: every feasible minimizing sequence converges to $\mathbf{0}$ in ℓ_1-norm whenever $\lambda > \lambda_{\max}$, and also at $\lambda = \lambda_{\max}$ if $\lambda_{\max} > 0$.

Finally, we demonstrate that the threshold is sharp. Let $\mathbf{e}_i$ denote the i-th standard basis vector in $\mathbb{R}^d$, and let $\mathbf{E}_{ij} := \mathbf{e}_i\mathbf{e}_j^\top$, the matrix with one in position (i,j) and zeros elsewhere. Suppose $0 \le \lambda < \lambda_{\max}$, and choose $i \ne j$ such that $\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right| = \lambda_{\max}$. For $t > 0$, define

$$
\mathbf{W}_t = t\ \operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\mathbf{E}_{ij}.
$$

Since $i \ne j$, we have $\mathbf{W}_t \odot \mathbf{W}_t = t^2\mathbf{E}_{ij}$ and $(\mathbf{W}_t \odot \mathbf{W}_t)^2 = t^4\mathbf{E}_{ij}^2 = t^4\mathbf{e}_i(\mathbf{e}_j^\top\mathbf{e}_i)\mathbf{e}_j^\top = \mathbf{0}$. Therefore

$$
e^{\mathbf{W}_t\odot\mathbf{W}_t} = \mathbf{I} + \mathbf{W}_t \odot \mathbf{W}_t = \mathbf{I} + t^2\mathbf{E}_{ij},
$$

and so

$$
h(\mathbf{W}_t) = \operatorname{tr}(e^{\mathbf{W}_t\odot\mathbf{W}_t}) - d = \operatorname{tr}(\mathbf{I} + t^2\mathbf{E}_{ij}) - d = 0.
$$

Thus $\mathbf{W}_t$ is feasible. Along this feasible ray, $\|\mathbf{W}_t\|_1 = t$, and

$$
\begin{aligned}
\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_t\right) &= t\ \operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{E}_{ij}\right) \\
&= t\ \operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ji} \\
&= t\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right| = t\lambda_{\max},
\end{aligned}
$$

where symmetry of $\mathbf{X}^\top\mathbf{X}/n$ is used, and

$$\begin{aligned}\frac{1}{2n}\|\mathbf{X}\mathbf{W}_t\|_F^2 &= \frac{1}{2n}\left\|t\ \operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\mathbf{X}\mathbf{e}_i\mathbf{e}_j^\top\right\|_F^2 \\ &= \frac{t^2}{2n}\|\mathbf{X}\mathbf{e}_i\mathbf{e}_j^\top\|_F^2 \\ &= \frac{t^2}{2n}\|\mathbf{X}\mathbf{e}_i\|_2^2\|\mathbf{e}_j\|_2^2 = \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii}.\end{aligned}$$

Hence

$$\begin{aligned}\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}_t\|_F^2+\lambda\|\mathbf{W}_t\|_1\right]-\left[\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{0}\|_F^2+\lambda\|\mathbf{0}\|_1\right] &= \frac{1}{2n}\|\mathbf{X}\mathbf{W}_t\|_F^2-\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_t\right)+\lambda\|\mathbf{W}_t\|_1 \\ &= \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii}-t\lambda_{\max}+\lambda t \\ &= \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii}-t(\lambda_{\max}-\lambda).\end{aligned}$$

Since $(\mathbf{X}^\top\mathbf{X}/n)_{ii}\geq 0$ and $\lambda_{\max}-\lambda>0$, the final expression is strictly negative for all sufficiently small $t>0$. Therefore no value below $\lambda_{\max}$ can guarantee that the zero matrix is globally minimizing, and the asserted threshold is exact. □

D Proof of Theorem 3.2

Proof. Let $\mathbf{W}^0=\mathbf{W}(\lambda_0)$ and $S=\operatorname{supp}(\mathbf{W}^0)$. Let $\mathcal{I}$ be the set of all $(i,j)\notin S$ such that $S\cup\{(i,j)\}$ is acyclic. Let $\mathbf{E}^{ij}$ denote the matrix satisfying $(\mathbf{E}^{ij})_{ab}=\mathbf{1}\{a=i,\ b=j\}$. Since $h(\mathbf{W}^0)=0$, the support S is acyclic. Thus any $\mathbf{W}$ with $\mathbf{W}_{S^c}=0$ satisfies $\operatorname{supp}(\mathbf{W})\subseteq S$, and consequently $h(\mathbf{W})=0$. For arbitrary $\mathbf{W},\mathbf{U}\in\mathbb{R}^{d\times d}$ and $t\in\mathbb{R}$, the quadratic loss satisfies the exact second-order identity

$$\begin{aligned}&\frac{1}{2n}\|\mathbf{X}-\mathbf{X}(\mathbf{W}+t\mathbf{U})\|_F^2-\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 \\ &\quad=\frac{1}{2n}\operatorname{tr}\big[(\mathbf{X}-\mathbf{X}\mathbf{W}-t\mathbf{X}\mathbf{U})^\top(\mathbf{X}-\mathbf{X}\mathbf{W}-t\mathbf{X}\mathbf{U})-(\mathbf{X}-\mathbf{X}\mathbf{W})^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\big] \\ &\quad=\frac{1}{2n}\operatorname{tr}\big[-t(\mathbf{X}\mathbf{U})^\top(\mathbf{X}-\mathbf{X}\mathbf{W})-t(\mathbf{X}-\mathbf{X}\mathbf{W})^\top\mathbf{X}\mathbf{U}+t^2(\mathbf{X}\mathbf{U})^\top\mathbf{X}\mathbf{U}\big] \\ &\quad=-\frac{t}{n}\operatorname{tr}\big[(\mathbf{X}\mathbf{U})^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\big]+\frac{t^2}{2n}\operatorname{tr}\big[(\mathbf{X}\mathbf{U})^\top\mathbf{X}\mathbf{U}\big] \\ &\quad=t\left\langle\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X}),\mathbf{U}\right\rangle_F+\frac{t^2}{2n}\|\mathbf{X}\mathbf{U}\|_F^2 \\ &\quad=t\sum_{a=1}^d\sum_{b=1}^d\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})\right]_{ab}\mathbf{U}_{ab}+\frac{t^2}{2n}\sum_{b=1}^d\sum_{r=1}^n\left(\sum_{a=1}^d\mathbf{X}_{ra}\mathbf{U}_{ab}\right)^2.\end{aligned}$$

For $(i,j)\in S$, the two matrices $\mathbf{W}^0+t\mathbf{E}^{ij}$ and $\mathbf{W}^0-t\mathbf{E}^{ij}$ are feasible for all sufficiently small $t>0$. Directional stationarity at $(\mathbf{W}^0,\lambda_0)$ gives

$$\begin{aligned}0&\leq\lim_{t\downarrow0}\frac{\frac{1}{2n}\|\mathbf{X}-\mathbf{X}(\mathbf{W}^0+t\mathbf{E}^{ij})\|_F^2+\lambda_0\|\mathbf{W}^0+t\mathbf{E}^{ij}\|_1-\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2-\lambda_0\|\mathbf{W}^0\|_1}{t} \\ &=\lim_{t\downarrow0}\left[\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\frac{t}{2n}\|\mathbf{X}\mathbf{E}^{ij}\|_F^2+\lambda_0\frac{|\mathbf{W}^0_{ij}+t|-|\mathbf{W}^0_{ij}|}{t}\right] \\ &=\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\lambda_0\operatorname{sgn}(\mathbf{W}^0_{ij}),\end{aligned}$$

and

$$
\begin{aligned}
0 &\le \lim_{t\downarrow 0} \frac{\frac{1}{2n}\|\mathbf{X}-\mathbf{X}(\mathbf{W}^0-t\mathbf{E}^{ij})\|_F^2+\lambda_0\|\mathbf{W}^0-t\mathbf{E}^{ij}\|_1-\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2-\lambda_0\|\mathbf{W}^0\|_1}{t} \\
&= \lim_{t\downarrow 0}\left[-\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\frac{t}{2n}\|\mathbf{X}\mathbf{E}^{ij}\|_F^2+\lambda_0\frac{|\mathbf{W}^0_{ij}-t|-|\mathbf{W}^0_{ij}|}{t}\right] \\
&= -\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}-\lambda_0\operatorname{sgn}(\mathbf{W}^0_{ij}).
\end{aligned}
$$

Thus

$$
\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})+\lambda_0\operatorname{sgn}(\mathbf{W}^0)\right]_{ij}=0, \qquad (i,j)\in S.
$$

For $(i,j)\in\mathcal{I}$, the matrices $\mathbf{W}^0+t\mathbf{E}^{ij}$ and $\mathbf{W}^0-t\mathbf{E}^{ij}$ are feasible for all sufficiently small $t>0$, and $\mathbf{W}^0_{ij}=0$. Hence

$$
\begin{aligned}
0 &\le \lim_{t\downarrow 0}\left[\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\frac{t}{2n}\|\mathbf{X}\mathbf{E}^{ij}\|_F^2+\lambda_0\frac{|t|-0}{t}\right] \\
&= \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\lambda_0,
\end{aligned}
$$

and

$$
\begin{aligned}
0 &\le \lim_{t\downarrow 0}\left[-\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\frac{t}{2n}\|\mathbf{X}\mathbf{E}^{ij}\|_F^2+\lambda_0\frac{|-t|-0}{t}\right] \\
&= -\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}+\lambda_0.
\end{aligned}
$$

Therefore

$$
\left|\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}\right|\le\lambda_0, \qquad (i,j)\in\mathcal{I}.
$$

Since $(\mathbf{W}^0,\lambda_0)$ is not a transition point, the inactive inequalities on $\mathcal{I}$ are strict, the active signs are nonzero, and the fixed-regime Jacobian is nonsingular. Equivalently,

$$
\gamma=\begin{cases}\min\limits_{(i,j)\in\mathcal{I}}\left\{\lambda_0-\left|\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0-\mathbf{X})\right]_{ij}\right|\right\}, & \mathcal{I}\neq\varnothing,\\ 1, & \mathcal{I}=\varnothing,\end{cases}\qquad \gamma>0,
$$

$$
\tau=\begin{cases}\min\limits_{(i,j)\in S}|\mathbf{W}^0_{ij}|, & S\neq\varnothing,\\ 1, & S=\varnothing,\end{cases}\qquad \tau>0,
$$

and

$$
\kappa=\begin{cases}\inf\limits_{\substack{\mathbf{U}_{S^c}=0\\ \|\mathbf{U}_S\|_F=1}}\left\|\left(\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\mathbf{U}\right]_{ij}\right)_{(i,j)\in S}\right\|_F, & S\neq\varnothing,\\ 1, & S=\varnothing,\end{cases}\qquad \kappa>0.
$$

Define the fixed-regime analytic stationarity map

$$
\mathbf{\Phi}(\mathbf{W}_S,\lambda)=\left(\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})+\lambda\operatorname{sgn}(\mathbf{W}^0)\right]_{ij}\right)_{(i,j)\in S}, \qquad \mathbf{W}_{S^c}=0.
$$

Then $\mathbf{\Phi}(\mathbf{W}_S^0, \lambda_0) = 0$, and, for $(i,j),(a,b),(c,d) \in S$,

$$\begin{aligned}
\frac{\partial \mathbf{\Phi}_{ij}}{\partial \mathbf{W}_{ab}}(\mathbf{W}_S, \lambda) &= \frac{\partial}{\partial \mathbf{W}_{ab}} \left\{ \frac{1}{n} \sum_{r=1}^{d} \mathbf{X}_{\cdot i}^\top \mathbf{X}_{\cdot r} \mathbf{W}_{rj} - \frac{1}{n} \mathbf{X}_{\cdot i}^\top \mathbf{X}_{\cdot j} + \lambda \operatorname{sgn}(\mathbf{W}_{ij}^0) \right\} \\
&= \frac{1}{n} \mathbf{X}_{\cdot i}^\top \mathbf{X}_{\cdot a} \mathbf{1}\{b = j\}, \\
\frac{\partial \mathbf{\Phi}_{ij}}{\partial \lambda}(\mathbf{W}_S, \lambda) &= \operatorname{sgn}(\mathbf{W}_{ij}^0), \\
\frac{\partial^2 \mathbf{\Phi}_{ij}}{\partial \mathbf{W}_{ab} \partial \mathbf{W}_{cd}}(\mathbf{W}_S, \lambda) &= 0, \\
\frac{\partial^2 \mathbf{\Phi}_{ij}}{\partial \lambda \partial \mathbf{W}_{ab}}(\mathbf{W}_S, \lambda) &= 0, \\
\frac{\partial^2 \mathbf{\Phi}_{ij}}{\partial \lambda^2}(\mathbf{W}_S, \lambda) &= 0.
\end{aligned}$$

Moreover, for every $\mathbf{U}_{S^c} = 0$,

$$\begin{aligned}
D_{\mathbf{W}_S} \mathbf{\Phi}(\mathbf{W}_S^0, \lambda_0)[\mathbf{U}_S] &= \left(\sum_{(a,b) \in S} \frac{\partial \mathbf{\Phi}_{ij}}{\partial \mathbf{W}_{ab}}(\mathbf{W}_S^0, \lambda_0) \mathbf{U}_{ab} \right)_{(i,j) \in S} \\
&= \left(\sum_{(a,b) \in S} \frac{1}{n} \mathbf{X}_{\cdot i}^\top \mathbf{X}_{\cdot a} \mathbf{1}\{b = j\} \mathbf{U}_{ab} \right)_{(i,j) \in S} \\
&= \left(\frac{1}{n} \sum_{\substack{a=1 \\ (a,j) \in S}}^{d} \mathbf{X}_{\cdot i}^\top \mathbf{X}_{\cdot a} \mathbf{U}_{aj} \right)_{(i,j) \in S} = \left(\left[\frac{1}{n} \mathbf{X}^\top \mathbf{X} \mathbf{U} \right]_{ij} \right)_{(i,j) \in S}.
\end{aligned}$$

The definition of κ gives the quantitative lower bound

$$\left\| D_{\mathbf{W}_S} \mathbf{\Phi}(\mathbf{W}_S^0, \lambda_0)[\mathbf{U}_S] \right\|_F \geq \kappa \|\mathbf{U}_S\|_F, \qquad \mathbf{U}_{S^c} = 0.$$

Since $\mathbf{\Phi}$ is real analytic and $D_{\mathbf{W}_S} \mathbf{\Phi}(\mathbf{W}_S^0, \lambda_0)$ is nonsingular, the analytic implicit function theorem yields $\varepsilon_1 > 0$, $r_1 > 0$, and a real-analytic map $\lambda \mapsto \mathbf{W}(\lambda)$ on $|\lambda - \lambda_0| \leq \varepsilon_1$, satisfying $\mathbf{W}(\lambda_0) = \mathbf{W}^0$, $\mathbf{W}_{S^c}(\lambda) = 0$, and $\mathbf{\Phi}(\mathbf{W}_S(\lambda), \lambda) = 0$. Moreover, if $|\lambda - \lambda_0| < \varepsilon_1$, $\widetilde{\mathbf{W}}_{S^c} = 0$, $\|\widetilde{\mathbf{W}} - \mathbf{W}^0\|_F < r_1$, and $\mathbf{\Phi}(\widetilde{\mathbf{W}}_S, \lambda) = 0$, then $\widetilde{\mathbf{W}} = \mathbf{W}(\lambda)$. Because $\mathbf{\Phi}$ is affine, the implicit branch satisfies the exact finite-difference identity

$$\begin{aligned}
0 &= \mathbf{\Phi}(\mathbf{W}_S(\lambda), \lambda) - \mathbf{\Phi}(\mathbf{W}_S^0, \lambda_0) \\
&= \left(\left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) + \lambda \operatorname{sgn}(\mathbf{W}^0) \right]_{ij} - \left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\mathbf{W}^0 - \mathbf{X}) + \lambda_0 \operatorname{sgn}(\mathbf{W}^0) \right]_{ij} \right)_{(i,j) \in S} \\
&= \left(\left[\frac{1}{n} \mathbf{X}^\top \mathbf{X} (\mathbf{W}(\lambda) - \mathbf{W}^0) \right]_{ij} + (\lambda - \lambda_0) \operatorname{sgn}(\mathbf{W}_{ij}^0) \right)_{(i,j) \in S}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\kappa \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_F &= \kappa \|\mathbf{W}_S(\lambda) - \mathbf{W}_S^0\|_F \\
&\leq \left\| \left(\left[\frac{1}{n} \mathbf{X}^\top \mathbf{X} (\mathbf{W}(\lambda) - \mathbf{W}^0) \right]_{ij} \right)_{(i,j) \in S} \right\|_F \\
&= |\lambda - \lambda_0| \left\| \left(\operatorname{sgn}(\mathbf{W}_{ij}^0) \right)_{(i,j) \in S} \right\|_F \\
&= |\lambda - \lambda_0| \sqrt{|S|}.
\end{aligned}$$

Set $M = \left\| \frac{1}{n}\mathbf{X}^\top \mathbf{X} \right\|_2$, and choose $0 < \varepsilon < \varepsilon_1$ such that

$$\varepsilon < \min\left\{ \frac{\lambda_0}{2}, \frac{\lambda_{\max} - \lambda_0}{2}, \frac{\kappa\tau}{8\max\{1, \sqrt{|S|}\}}, \frac{\gamma}{8\{1 + M\kappa^{-1}\max\{1, \sqrt{|S|}\}\}}, \frac{r_1\kappa}{2\max\{1, \sqrt{|S|}\}} \right\}.$$

Then, whenever $|\lambda - \lambda_0| < \varepsilon$,

$$\begin{aligned} 0 < \lambda < \lambda_{\max}, \\ \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_F \le \kappa^{-1}\sqrt{|S|}\,|\lambda - \lambda_0| < \frac{\tau}{8}, \\ \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_F < \frac{r_1}{2}. \end{aligned}$$

For every $(i,j) \in S$,

$$\begin{aligned} |\mathbf{W}_{ij}(\lambda)| &= |\mathbf{W}^0_{ij} + \mathbf{W}_{ij}(\lambda) - \mathbf{W}^0_{ij}| \\ &\ge |\mathbf{W}^0_{ij}| - |\mathbf{W}_{ij}(\lambda) - \mathbf{W}^0_{ij}| \\ &\ge |\mathbf{W}^0_{ij}| - \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_{\max} \\ &\ge |\mathbf{W}^0_{ij}| - \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_F \\ &\ge \tau - \kappa^{-1}\sqrt{|S|}\,|\lambda - \lambda_0| \\ &> \tau - \frac{\tau}{8} = \frac{7\tau}{8} > 0. \end{aligned}$$

Hence $\mathrm{sgn}(\mathbf{W}_{ij}(\lambda)) = \mathrm{sgn}(\mathbf{W}^0_{ij})$ for every $(i,j) \in S$. Since $\mathbf{W}_{S^c}(\lambda) = 0$, $\mathrm{supp}(\mathbf{W}(\lambda)) = S$. Since S is acyclic, $h(\mathbf{W}(\lambda)) = 0$ for $|\lambda - \lambda_0| < \varepsilon$. For every $(i,j) \in \mathcal{I}$, the inactive margin along the branch satisfies

$$\begin{aligned} &\lambda - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \right| \\ &= \lambda - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) + \frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda) - \mathbf{W}^0) \right]_{ij} \right| \\ &\ge \lambda_0 - |\lambda - \lambda_0| - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) + \frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda) - \mathbf{W}^0) \right]_{ij} \right| \\ &\ge \lambda_0 - |\lambda - \lambda_0| - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) \right]_{ij} \right| - \left| \left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda) - \mathbf{W}^0) \right]_{ij} \right| \\ &\ge \lambda_0 - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) \right]_{ij} \right| - |\lambda - \lambda_0| - \left\| \frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda) - \mathbf{W}^0) \right\|_F \\ &\ge \lambda_0 - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) \right]_{ij} \right| - |\lambda - \lambda_0| - \left\| \frac{1}{n}\mathbf{X}^\top\mathbf{X} \right\|_2 \|\mathbf{W}(\lambda) - \mathbf{W}^0\|_F \\ &\ge \lambda_0 - \left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}^0 - \mathbf{X}) \right]_{ij} \right| - |\lambda - \lambda_0| - M\kappa^{-1}\sqrt{|S|}\,|\lambda - \lambda_0| \\ &\ge \gamma - \{1 + M\kappa^{-1}\max\{1, \sqrt{|S|}\}\}|\lambda - \lambda_0| \\ &> \gamma - \frac{\gamma}{8} = \frac{7\gamma}{8} > 0. \end{aligned}$$

Equivalently, $\left| \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \right| < \lambda$ for every $(i,j) \in \mathcal{I}$ and $|\lambda - \lambda_0| < \varepsilon$. Fix $|\lambda - \lambda_0| < \varepsilon$. Let $\mathbf{U} \in \mathbb{R}^{d \times d}$ be any one-sided feasible direction at $\mathbf{W}(\lambda)$, namely, $h(\mathbf{W}(\lambda) + t\mathbf{U}) = 0$ for all

sufficiently small $t > 0$. For every $0 < t < \dfrac{7\tau}{16(1+\|\mathbf{U}\|_{\max})}$ and every $(a,b) \in S$,

$$\begin{aligned}
|\mathbf{W}_{ab}(\lambda) + t\mathbf{U}_{ab}| &\geq |\mathbf{W}_{ab}(\lambda)| - t|\mathbf{U}_{ab}| \\
&> \frac{7\tau}{8} - \frac{7\tau}{16(1+\|\mathbf{U}\|_{\max})}\|\mathbf{U}\|_{\max} \\
&\geq \frac{7\tau}{8} - \frac{7\tau}{16} = \frac{7\tau}{16} > 0.
\end{aligned}$$

Therefore

$$\begin{gathered}
S \subseteq \operatorname{supp}(\mathbf{W}(\lambda) + t\mathbf{U}), \\
\{(i,j) \notin S : \mathbf{U}_{ij} \neq 0\} \subseteq \operatorname{supp}(\mathbf{W}(\lambda) + t\mathbf{U}).
\end{gathered}$$

Since $h(\mathbf{W}(\lambda) + t\mathbf{U}) = 0$, the support of $\mathbf{W}(\lambda) + t\mathbf{U}$ is acyclic. Hence, if $(i,j) \notin S$ and $\mathbf{U}_{ij} \neq 0$, then $S \cup \{(i,j)\}$ is acyclic, so $(i,j) \in \mathcal{I}$. For the loss part of the one-sided directional derivative,

$$\begin{aligned}
&\lim_{t\downarrow 0} \frac{\frac{1}{2n}\|\mathbf{X} - \mathbf{X}(\mathbf{W}(\lambda) + t\mathbf{U})\|_F^2 - \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda)\|_F^2}{t} \\
&= \lim_{t\downarrow 0} \left[\left\langle \frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}), \mathbf{U} \right\rangle_F + \frac{t}{2n}\|\mathbf{X}\mathbf{U}\|_F^2 \right] \\
&= \left\langle \frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}), \mathbf{U} \right\rangle_F \\
&= \sum_{(i,j)\in S} \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij} + \sum_{(i,j)\notin S} \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij}.
\end{aligned}$$

For the ℓ_1-part, using the active sign preservation and $\mathbf{W}_{S^c}(\lambda) = 0$,

$$\begin{aligned}
&\lim_{t\downarrow 0} \frac{\|\mathbf{W}(\lambda) + t\mathbf{U}\|_1 - \|\mathbf{W}(\lambda)\|_1}{t} \\
&= \lim_{t\downarrow 0} \frac{\sum_{(i,j)\in S} \{|\mathbf{W}_{ij}(\lambda) + t\mathbf{U}_{ij}| - |\mathbf{W}_{ij}(\lambda)|\} + \sum_{(i,j)\notin S} |t\mathbf{U}_{ij}|}{t} \\
&= \sum_{(i,j)\in S} \lim_{t\downarrow 0} \frac{|\mathbf{W}_{ij}(\lambda) + t\mathbf{U}_{ij}| - |\mathbf{W}_{ij}(\lambda)|}{t} + \sum_{(i,j)\notin S} |\mathbf{U}_{ij}| \\
&= \sum_{(i,j)\in S} \operatorname{sgn}(\mathbf{W}_{ij}(\lambda))\mathbf{U}_{ij} + \sum_{(i,j)\notin S} |\mathbf{U}_{ij}| \\
&= \sum_{(i,j)\in S} \operatorname{sgn}(\mathbf{W}^0_{ij})\mathbf{U}_{ij} + \sum_{(i,j)\notin S} |\mathbf{U}_{ij}|.
\end{aligned}$$

Combining the two exact directional derivative identities gives

$$\begin{aligned}
&\lim_{t\downarrow 0} \frac{\frac{1}{2n}\|\mathbf{X} - \mathbf{X}(\mathbf{W}(\lambda) + t\mathbf{U})\|_F^2 + \lambda\|\mathbf{W}(\lambda) + t\mathbf{U}\|_1 - \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda)\|_F^2 - \lambda\|\mathbf{W}(\lambda)\|_1}{t} \\
&= \sum_{(i,j)\in S} \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij} + \sum_{(i,j)\notin S} \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij} \\
&\quad + \lambda \sum_{(i,j)\in S} \operatorname{sgn}(\mathbf{W}^0_{ij})\mathbf{U}_{ij} + \lambda \sum_{(i,j)\notin S} |\mathbf{U}_{ij}| \\
&= \sum_{(i,j)\in S} \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) + \lambda \operatorname{sgn}(\mathbf{W}^0) \right]_{ij} \mathbf{U}_{ij} + \sum_{(i,j)\notin S} \left\{ \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij} + \lambda|\mathbf{U}_{ij}| \right\} \\
&= \sum_{(i,j)\notin S} \left\{ \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \mathbf{U}_{ij} + \lambda|\mathbf{U}_{ij}| \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{(i,j)\notin S\\ \mathbf{U}_{ij}\neq 0}} \left\{ \left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda)-\mathbf{X})\right]_{ij}\mathbf{U}_{ij} + \lambda|\mathbf{U}_{ij}| \right\} \\
&\geq \sum_{\substack{(i,j)\notin S\\ \mathbf{U}_{ij}\neq 0}} \left\{ -\left|\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda)-\mathbf{X})\right]_{ij}\right| |\mathbf{U}_{ij}| + \lambda|\mathbf{U}_{ij}| \right\} \\
&= \sum_{\substack{(i,j)\notin S\\ \mathbf{U}_{ij}\neq 0}} \left\{ \lambda - \left|\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda)-\mathbf{X})\right]_{ij}\right| \right\} |\mathbf{U}_{ij}| \\
&= \sum_{\substack{(i,j)\notin S\\ \mathbf{U}_{ij}\neq 0\\ (i,j)\in\mathcal{I}}} \left\{ \lambda - \left|\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{W}(\lambda)-\mathbf{X})\right]_{ij}\right| \right\} |\mathbf{U}_{ij}| \\
&\geq \frac{7\gamma}{8} \sum_{\substack{(i,j)\notin S\\ \mathbf{U}_{ij}\neq 0\\ (i,j)\in\mathcal{I}}} |\mathbf{U}_{ij}| \geq 0.
\end{aligned}
$$

Hence every feasible one-sided direction at $\mathbf{W}(\lambda)$ has nonnegative one-sided directional derivative, and therefore $\mathbf{W}(\lambda)$ is a stationary point of (1) for every $\lambda \in (\lambda_0-\varepsilon, \lambda_0+\varepsilon)$. The map $\lambda \mapsto \mathbf{W}(\lambda)$ is real analytic by the analytic implicit function theorem. Therefore, for every integer $m \geq 0$, $\dfrac{d^m}{d\lambda^m}\mathbf{W}(\lambda)$ exists and is continuous on $(\lambda_0-\varepsilon, \lambda_0+\varepsilon)$, and $\lambda \mapsto \mathbf{W}(\lambda) \in C^\omega((\lambda_0-\varepsilon, \lambda_0+\varepsilon); \mathbb{R}^{d\times d}) \subset C^\infty((\lambda_0-\varepsilon, \lambda_0+\varepsilon); \mathbb{R}^{d\times d})$.

For the continuation assertion, let $J \subset (0, \lambda_{\max})$ be a connected interval on which the branch contains no transition point. On any compact subinterval of J, the support, active signs, acyclic stratum, inactive margins, and fixed-regime nonsingularity remain stable. For any $\lambda, \lambda' \in J$ lying in a common fixed regime,

$$
\begin{aligned}
0 &= \mathbf{\Phi}(\mathbf{W}_S(\lambda), \lambda) - \mathbf{\Phi}(\mathbf{W}_S(\lambda'), \lambda') \\
&= \left(\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda)-\mathbf{W}(\lambda'))\right]_{ij} + (\lambda-\lambda')\operatorname{sgn}(\mathbf{W}^0_{ij}) \right)_{(i,j)\in S},
\end{aligned}
$$

and hence

$$
\begin{aligned}
\kappa\|\mathbf{W}(\lambda)-\mathbf{W}(\lambda')\|_F &= \kappa\|\mathbf{W}_S(\lambda)-\mathbf{W}_S(\lambda')\|_F \\
&\leq \left\| \left(\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{W}(\lambda)-\mathbf{W}(\lambda'))\right]_{ij} \right)_{(i,j)\in S} \right\|_F \\
&= |\lambda-\lambda'| \left\| \left(\operatorname{sgn}(\mathbf{W}^0_{ij})\right)_{(i,j)\in S} \right\|_F \\
&= |\lambda-\lambda'|\sqrt{|S|}.
\end{aligned}
$$

Thus $\|\mathbf{W}(\lambda)-\mathbf{W}(\lambda')\|_F \leq \kappa^{-1}\sqrt{|S|}\,|\lambda-\lambda'|$. Let $J = (\underline{\lambda}, \overline{\lambda})$ with $\overline{\lambda} < \lambda_{\max}$. The preceding estimate gives

$$
\begin{aligned}
\sup_{\lambda,\lambda'\in(\overline{\lambda}-\eta,\overline{\lambda})} \|\mathbf{W}(\lambda)-\mathbf{W}(\lambda')\|_F &\leq \kappa^{-1}\sqrt{|S|} \sup_{\lambda,\lambda'\in(\overline{\lambda}-\eta,\overline{\lambda})} |\lambda-\lambda'| \\
&\leq \kappa^{-1}\sqrt{|S|}\,\eta.
\end{aligned}
$$

This upper bound converges to zero as $\eta \downarrow 0$. Since $\mathbb{R}^{d\times d}$ is complete, there exists $\overline{\mathbf{W}} \in \mathbb{R}^{d\times d}$ such that $\lim_{\lambda\uparrow\overline{\lambda}} \mathbf{W}(\lambda) = \overline{\mathbf{W}}$. Moreover, $\overline{\mathbf{W}}_{S^c} = \lim_{\lambda\uparrow\overline{\lambda}} \mathbf{W}_{S^c}(\lambda) = 0$, and, by continuity of h,

$$
h(\overline{\mathbf{W}}) = h\left(\lim_{\lambda\uparrow\overline{\lambda}} \mathbf{W}(\lambda)\right) = \lim_{\lambda\uparrow\overline{\lambda}} h(\mathbf{W}(\lambda)) = 0.
$$

Also,

$$\mathbf{\Phi}(\overline{\mathbf{W}}_S, \overline{\lambda}) = \left(\left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\overline{\mathbf{W}} - \mathbf{X}) + \overline{\lambda} \operatorname{sgn}(\mathbf{W}^0) \right]_{ij} \right)_{(i,j)\in S}$$
$$= \lim_{\lambda \uparrow \overline{\lambda}} \left(\left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) + \lambda \operatorname{sgn}(\mathbf{W}^0) \right]_{ij} \right)_{(i,j)\in S} = \lim_{\lambda \uparrow \overline{\lambda}} \mathbf{\Phi}(\mathbf{W}_S(\lambda), \lambda) = 0.$$

For every $(i,j) \in \mathcal{I}$,

$$\overline{\lambda} - \left| \left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\overline{\mathbf{W}} - \mathbf{X}) \right]_{ij} \right| = \lim_{\lambda \uparrow \overline{\lambda}} \left\{ \lambda - \left| \left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\mathbf{W}(\lambda) - \mathbf{X}) \right]_{ij} \right| \right\}.$$

If

$$\min_{(i,j)\in S} |\overline{\mathbf{W}}_{ij}| > 0,$$

$$\min_{(i,j)\in \mathcal{I}} \left\{ \overline{\lambda} - \left| \left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\overline{\mathbf{W}} - \mathbf{X}) \right]_{ij} \right| \right\} > 0,$$

$$\inf_{\substack{\mathbf{U}_{S^c}=0 \\ \|\mathbf{U}_S\|_F=1}} \left\| \left(\left[\frac{1}{n} \mathbf{X}^\top \mathbf{X}\mathbf{U} \right]_{ij} \right)_{(i,j)\in S} \right\|_F > 0,$$

and the acyclic stratum does not change at $\overline{\mathbf{W}}$, then $(\overline{\mathbf{W}}, \overline{\lambda})$ is not a transition point. Applying the local construction already proved, with $(\mathbf{W}^0, \lambda_0)$ replaced by $(\overline{\mathbf{W}}, \overline{\lambda})$, gives $\eta > 0$, a corresponding uniqueness radius $\bar{r}_1 > 0$, and a real-analytic stationary branch $\lambda \mapsto \widetilde{\mathbf{W}}(\lambda)$ on $|\lambda - \overline{\lambda}| < \eta$, satisfying $\widetilde{\mathbf{W}}(\overline{\lambda}) = \overline{\mathbf{W}}$, $\widetilde{\mathbf{W}}_{S^c}(\lambda) = 0$, and $\mathbf{\Phi}(\widetilde{\mathbf{W}}_S(\lambda), \lambda) = 0$. For $\lambda \in (\overline{\lambda} - \eta, \overline{\lambda}) \cap J$, both $\mathbf{W}(\lambda)$ and $\widetilde{\mathbf{W}}(\lambda)$ solve the same fixed-regime equation. Moreover,

$$\lim_{\lambda \uparrow \overline{\lambda}} \|\mathbf{W}(\lambda) - \overline{\mathbf{W}}\|_F = 0,$$
$$\lim_{\lambda \uparrow \overline{\lambda}} \|\widetilde{\mathbf{W}}(\lambda) - \overline{\mathbf{W}}\|_F = 0.$$

Thus, for sufficiently small $\eta' > 0$ and every $\lambda \in (\overline{\lambda} - \eta', \overline{\lambda})$, both $\|\mathbf{W}(\lambda) - \overline{\mathbf{W}}\|_F < \bar{r}_1$ and $\|\widetilde{\mathbf{W}}(\lambda) - \overline{\mathbf{W}}\|_F < \bar{r}_1$. Local uniqueness in the implicit-function theorem yields $\mathbf{W}(\lambda) = \widetilde{\mathbf{W}}(\lambda)$ for $\lambda \in (\overline{\lambda} - \eta', \overline{\lambda})$. Equivalently, the piecewise-defined map

$$\lambda \mapsto \begin{cases} \mathbf{W}(\lambda), & \lambda \in J, \\ \widetilde{\mathbf{W}}(\lambda), & \lambda \in (\overline{\lambda} - \eta', \overline{\lambda} + \eta') \end{cases}$$

is a real-analytic stationary extension beyond $\overline{\lambda}$, contradicting maximality of J as a transition-free continuation interval. Hence, if $\overline{\lambda} < \lambda_{\max}$, maximality forces at least one of the following transition conditions:

$$\min_{(i,j)\in S} |\overline{\mathbf{W}}_{ij}| = 0,$$

$$\min_{(i,j)\in \mathcal{I}} \left\{ \overline{\lambda} - \left| \left[\frac{1}{n} \mathbf{X}^\top (\mathbf{X}\overline{\mathbf{W}} - \mathbf{X}) \right]_{ij} \right| \right\} = 0,$$

$$\inf_{\substack{\mathbf{U}_{S^c}=0 \\ \|\mathbf{U}_S\|_F=1}} \left\| \left(\left[\frac{1}{n} \mathbf{X}^\top \mathbf{X}\mathbf{U} \right]_{ij} \right)_{(i,j)\in S} \right\|_F = 0.$$

The remaining possibility is a change of acyclic stratum. The same argument with $\lambda \downarrow \underline{\lambda}$ gives the corresponding statement at the left endpoint. These alternatives are precisely the transition alternatives. Therefore the local regularization path is C^∞ smooth and admits analytic continuation up to the nearest transition point. □

E Proof of Theorem 3.3

Proof. Let $I \subset (0, \lambda_{\max})$ be an open interval containing λ_0 on which $\lambda \mapsto \mathbf{W}(\lambda)$ is the analytic branch supplied by Theorem 3.2. Since $(\mathbf{W}(\lambda_0), \lambda_0)$ is not a transition point, shrink I, if necessary, so that the active set, the active signs, and one topological order remain fixed on I. Denote this order by $p_1, \dots, p_d$. Thus, for every $\lambda \in I$,

$$\begin{aligned} \mathcal{A} &= \{(i,j) : \mathbf{W}_{ij}(\lambda) \neq 0\}, \\ \operatorname{sign}(\mathbf{W}_{ij}(\lambda)) &= \operatorname{sign}(\mathbf{W}_{ij}(\lambda_0)), \qquad (i,j) \in \mathcal{A}, \\ \mathbf{W}_{p_a p_b}(\lambda) &= 0, \qquad 1 \le b \le a \le d. \end{aligned}$$

For each column j, write $\mathcal{A}_j = \{i : (i,j) \in \mathcal{A}\}$. Let $\mathbf{D} \in \mathbb{R}^{d \times d}$ satisfy $\mathbf{D}_{p_a p_b} = 0$ for $a \ge b$. For $\lambda \in I$ and $t \in \mathbb{R}$, define $\mathbf{B}_{\lambda,t} = (\mathbf{W}(\lambda) + t\mathbf{D}) \odot (\mathbf{W}(\lambda) + t\mathbf{D})$. Since $\mathbf{W}_{p_r p_s}(\lambda) + t\mathbf{D}_{p_r p_s} = 0$ whenever $r \ge s$, the matrix $\mathbf{B}_{\lambda,t}$ is strictly upper triangular in the order $p_1, \dots, p_d$. Hence $[\mathbf{B}_{\lambda,t}^m]_{p_a p_a} = 0$ for all $m \ge 1$ and all $a \in \{1, \dots, d\}$. Therefore

$$\begin{aligned} h(\mathbf{W}(\lambda) + t\mathbf{D}) &= \operatorname{tr}(\exp(\mathbf{B}_{\lambda,t})) - d \\ &= \sum_{a=1}^{d} \left[1 + \sum_{m=1}^{\infty} \frac{1}{m!} [\mathbf{B}_{\lambda,t}^m]_{p_a p_a} \right] - d = 0. \end{aligned}$$

Thus, on this non-transition regime, the exact local h-feasible chart is the linear topological chart $\mathbf{W}_{p_a p_b} = 0$ for $a \ge b$. For $t > 0$ sufficiently small such that $|t\mathbf{D}_{ij}| < |\mathbf{W}_{ij}(\lambda)|$ for every $(i,j) \in \mathcal{A}$ with $\mathbf{D}_{ij} \neq 0$, define

$$\Delta_t(\mathbf{D}; \lambda) = \frac{1}{t} \left[\frac{1}{2n} \left\| \mathbf{X} - \mathbf{X}\big(\mathbf{W}(\lambda) + t\mathbf{D}\big) \right\|_F^2 + \lambda \left\| \mathbf{W}(\lambda) + t\mathbf{D} \right\|_1 - \frac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda)\|_F^2 - \lambda \|\mathbf{W}(\lambda)\|_1 \right].$$

The exact finite-increment identity along this feasible chart is

$$\begin{aligned} \Delta_t(\mathbf{D}; \lambda) &= \frac{1}{2nt} \left[\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda) - t\mathbf{X}\mathbf{D}\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda)\|_F^2 \right] \\ &\quad + \frac{\lambda}{t} \left[\sum_{(i,j) \in \mathcal{A}} \left(|\mathbf{W}_{ij}(\lambda) + t\mathbf{D}_{ij}| - |\mathbf{W}_{ij}(\lambda)| \right) + \sum_{\substack{(i,j) \notin \mathcal{A} \\ \mathbf{D}_{ij} \neq 0}} |t\mathbf{D}_{ij}| \right] \\ &= \sum_{j=1}^{d} \sum_{i=1}^{d} \mathbf{D}_{ij} \frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) + \lambda \sum_{(i,j) \in \mathcal{A}} \operatorname{sign}(\mathbf{W}_{ij}(\lambda)) \mathbf{D}_{ij} + \lambda \sum_{\substack{(i,j) \notin \mathcal{A} \\ \mathbf{D}_{ij} \neq 0}} |\mathbf{D}_{ij}| + \frac{t}{2n} \|\mathbf{X}\mathbf{D}\|_F^2 \\ &= \sum_{(i,j) \in \mathcal{A}} \mathbf{D}_{ij} \left[\frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) + \lambda \operatorname{sign}(\mathbf{W}_{ij}(\lambda)) \right] \\ &\quad + \sum_{\substack{(i,j) \notin \mathcal{A} \\ \mathbf{D}_{ij} \neq 0}} \left[\mathbf{D}_{ij} \frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) + \lambda |\mathbf{D}_{ij}| \right] + \frac{t}{2n} \|\mathbf{X}\mathbf{D}\|_F^2. \end{aligned}$$

The remainder satisfies $0 \le \frac{t}{2n} \|\mathbf{X}\mathbf{D}\|_F^2 \le \frac{t}{2n} \|\mathbf{X}\|_{\mathrm{op}}^2 \|\mathbf{D}\|_F^2$, and hence vanishes as $t \downarrow 0$. Since $\mathbf{W}(\lambda)$ is stationary for the constrained problem on the exact feasible chart, every feasible chart direction $\mathbf{D}$ satisfies $0 \le \liminf_{t \downarrow 0} \Delta_t(\mathbf{D}; \lambda)$. Taking $\mathbf{D} = \mathbf{E}^{ij}$ and $\mathbf{D} = -\mathbf{E}^{ij}$, where $\mathbf{E}^{ij}$ has (i,j)-entry equal to 1 and all remaining entries equal to 0, gives, for every $(i,j) \in \mathcal{A}$,

$$\begin{aligned} 0 &\le \frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) + \lambda \operatorname{sign}(\mathbf{W}_{ij}(\lambda)), \\ 0 &\le -\frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) - \lambda \operatorname{sign}(\mathbf{W}_{ij}(\lambda)). \end{aligned}$$

Hence

$$\frac{1}{n} \mathbf{X}_{:,i}^\top \big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j} \big) + \lambda \operatorname{sign}(\mathbf{W}_{ij}(\lambda)) = 0, \qquad (i,j) \in \mathcal{A}.$$

For an inactive coordinate $(i,j) \notin \mathcal{A}$ compatible with the fixed order, namely $i = p_a$, $j = p_b$, and $a < b$, the same two one-sided perturbations give

$$0 \le \frac{1}{n}\mathbf{X}_{:,i}^\top\big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j}\big) + \lambda,$$
$$0 \le -\frac{1}{n}\mathbf{X}_{:,i}^\top\big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j}\big) + \lambda.$$

Equivalently,

$$\left|\frac{1}{n}\mathbf{X}_{:,i}^\top\big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j}\big)\right| \le \lambda, \qquad (i,j) \notin \mathcal{A}, \quad i = p_a, \quad j = p_b, \quad a < b.$$

Because $\mathbf{W}_{kj}(\lambda) = 0$ for $k \notin \mathcal{A}_j$, the active equalities become

$$\begin{aligned} 0 &= \frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right) + \lambda\,\mathrm{sign}(\mathbf{W}_{ij}(\lambda)) \\ &= \sum_{k\in\mathcal{A}_j}\frac{1}{n}\mathbf{X}_{:,i}^\top\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \frac{1}{n}\mathbf{X}_{:,i}^\top\mathbf{X}_{:,j} + \lambda\,\mathrm{sign}(\mathbf{W}_{ij}(\lambda)), \qquad i \in \mathcal{A}_j. \end{aligned}$$

On the non-transition regime, the feasible inactive inequalities are strict:

$$\left|\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right| < \lambda, \qquad (i,j) \notin \mathcal{A}, \quad i = p_a, \quad j = p_b, \quad a < b.$$

Conversely, suppose that the active equalities and feasible inactive inequalities in the previous two displays hold on the fixed topological chart. Then, for every $\mathbf{D}$ satisfying $\mathbf{D}_{p_a p_b} = 0$ for $a \ge b$, the finite-increment identity yields

$$\begin{aligned} \liminf_{t\downarrow 0}\Delta_t(\mathbf{D};\lambda) &= \sum_{(i,j)\in\mathcal{A}}\mathbf{D}_{ij}\left[\frac{1}{n}\mathbf{X}_{:,i}^\top\big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j}\big) + \lambda\,\mathrm{sign}(\mathbf{W}_{ij}(\lambda))\right] \\ &\quad + \sum_{\substack{(i,j)\notin\mathcal{A}\\ \mathbf{D}_{ij}\neq 0}}\left[\mathbf{D}_{ij}\frac{1}{n}\mathbf{X}_{:,i}^\top\big(\mathbf{X}\mathbf{W}_{:,j}(\lambda) - \mathbf{X}_{:,j}\big) + \lambda|\mathbf{D}_{ij}|\right] \\ &\ge \sum_{\substack{(i,j)\notin\mathcal{A}\\ \mathbf{D}_{ij}\neq 0}}|\mathbf{D}_{ij}|\left[\lambda - \left|\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right|\right] \ge 0. \end{aligned}$$

Together with the exact feasibility $h(\mathbf{W}(\lambda) + t\mathbf{D}) = 0$, this proves the local equivalence between constrained NOTEARS stationarity on the fixed feasible chart and the stratum-restricted stationary system

$$\begin{cases} \mathbf{W}_{ij}(\lambda) = 0, & (i,j) \notin \mathcal{A}, \\ \displaystyle\sum_{k\in\mathcal{A}_j}\frac{1}{n}\mathbf{X}_{:,i}^\top\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \frac{1}{n}\mathbf{X}_{:,i}^\top\mathbf{X}_{:,j} + \lambda\,\mathrm{sign}(\mathbf{W}_{ij}(\lambda)) = 0, & i \in \mathcal{A}_j, \quad j = 1,\dots,d, \\ \left|\dfrac{1}{n}\mathbf{X}_{:,i}^\top\left(\displaystyle\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right| < \lambda, & (i,j) \notin \mathcal{A}, \quad i = p_a, \quad j = p_b, \quad a < b. \end{cases}$$

The path equation follows by comparing the active stationary system at two nearby regularization values. Fix $j \in \{1,\dots,d\}$. If $\mathcal{A}_j = \varnothing$, there is no active equation for this column. Otherwise, for every $i \in \mathcal{A}_j$, every $\lambda \in I$, and every $\delta \neq 0$ with $\lambda + \delta \in I$, subtracting the active equation at λ from

the active equation at $\lambda + \delta$ gives

$$\begin{aligned}0 &= \left[\sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda+\delta) - \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + (\lambda+\delta)\operatorname{sign}(\mathbf{W}_{ij}(\lambda+\delta))\right] \\ &\quad - \left[\sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda\operatorname{sign}(\mathbf{W}_{ij}(\lambda))\right] \\ &= \sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\left[\mathbf{W}_{kj}(\lambda+\delta) - \mathbf{W}_{kj}(\lambda)\right] + \delta\operatorname{sign}(\mathbf{W}_{ij}(\lambda)).\end{aligned}$$

The active signs are fixed on I, so $\operatorname{sign}(\mathbf{W}_{ij}(\lambda+\delta)) = \operatorname{sign}(\mathbf{W}_{ij}(\lambda)) = \operatorname{sign}(\mathbf{W}_{ij}(\lambda_0))$ for $i \in \mathcal{A}_j$. Dividing by δ gives the exact finite-difference system

$$\sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\frac{\mathbf{W}_{kj}(\lambda+\delta) - \mathbf{W}_{kj}(\lambda)}{\delta} = -\operatorname{sign}(\mathbf{W}_{ij}(\lambda)), \qquad i \in \mathcal{A}_j.$$

By analyticity of $\lambda \mapsto \mathbf{W}(\lambda)$, each scalar function $\lambda \mapsto \mathbf{W}_{kj}(\lambda)$ satisfies

$$\frac{\mathbf{W}_{kj}(\lambda+\delta) - \mathbf{W}_{kj}(\lambda)}{\delta} = \frac{d\mathbf{W}_{kj}(\lambda)}{d\lambda} + \delta\int_0^1 (1-s)\frac{d^2\mathbf{W}_{kj}(\lambda+s\delta)}{d\lambda^2}\,ds, \qquad k \in \mathcal{A}_j.$$

Consequently,

$$\begin{aligned}&\left|\sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\frac{\mathbf{W}_{kj}(\lambda+\delta) - \mathbf{W}_{kj}(\lambda)}{\delta} - \sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\frac{d\mathbf{W}_{kj}(\lambda)}{d\lambda}\right| \\ &\le |\delta| \sum_{k\in\mathcal{A}_j} \left|\frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\right| \int_0^1 (1-s)\left|\frac{d^2\mathbf{W}_{kj}(\lambda+s\delta)}{d\lambda^2}\right| ds \\ &\le \frac{|\delta|}{2} \sum_{k\in\mathcal{A}_j} \left|\frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\right| \sup_{\eta:\ |\eta-\lambda|\le|\delta|} \left|\frac{d^2\mathbf{W}_{kj}(\eta)}{d\lambda^2}\right|.\end{aligned}$$

The right-hand side converges to zero as $\delta \to 0$. Passing to the limit in the finite-difference system yields

$$\sum_{k\in\mathcal{A}_j} \frac{1}{n}\mathbf{X}_{:,i}^\top \mathbf{X}_{:,k}\frac{d\mathbf{W}_{kj}(\lambda)}{d\lambda} = -\operatorname{sign}(\mathbf{W}_{ij}(\lambda)), \qquad i \in \mathcal{A}_j.$$

Stacking these equations over $i \in \mathcal{A}_j$ gives

$$\left(\frac{1}{n}\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}\right)\frac{d\mathbf{W}_{\mathcal{A}_j,j}(\lambda)}{d\lambda} = -\operatorname{sign}\big(\mathbf{W}_{\mathcal{A}_j,j}(\lambda)\big), \qquad j = 1,\dots,d.$$

It remains to verify persistence of inactive entries on the same local regime. If $(i,j) \notin \mathcal{A}$ is incompatible with the fixed topological order, then $i = p_a$, $j = p_b$ for some $a \ge b$, and the chart equation gives $\mathbf{W}_{ij}(\lambda) = \mathbf{W}_{p_a p_b}(\lambda) = 0$ on I. If $(i,j) \notin \mathcal{A}$ is compatible with the fixed topological order, then $i = p_a$, $j = p_b$ for some $a < b$. For every compact interval $[\lambda_-, \lambda_+] \subset I$, strict inactive feasibility and continuity give

$$\inf_{\lambda\in[\lambda_-,\lambda_+]} \left\{\lambda - \left|\frac{1}{n}\mathbf{X}_{:,i}^\top \left(\sum_{k\in\mathcal{A}_j} \mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right|\right\} > 0.$$

For every $t \neq 0$ and every $\lambda \in [\lambda_-, \lambda_+]$,

$$\begin{aligned}
&\frac{1}{2n}\left\|\mathbf{X}_{:,j} - \sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - t\mathbf{X}_{:,i}\right\|_2^2 + \lambda|t| - \frac{1}{2n}\left\|\mathbf{X}_{:,j} - \sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda)\right\|_2^2 \\
&= -\frac{t}{n}\mathbf{X}_{:,i}^\top\left(\mathbf{X}_{:,j} - \sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda)\right) + \frac{t^2}{2n}\|\mathbf{X}_{:,i}\|_2^2 + \lambda|t| \\
&\geq |t|\left[\lambda - \left|\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right|\right] + \frac{t^2}{2n}\|\mathbf{X}_{:,i}\|_2^2 \\
&\geq |t| \inf_{\eta\in[\lambda_-,\lambda_+]}\left\{\eta - \left|\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\eta) - \mathbf{X}_{:,j}\right)\right|\right\} + \frac{t^2}{2n}\|\mathbf{X}_{:,i}\|_2^2 > 0.
\end{aligned}$$

Thus a compatible inactive coordinate cannot leave zero while the strict inactive margin remains positive. Equivalently, if such an entry became active inside I, then there would exist $\lambda_r \to \bar{\lambda} \in I$ and $s \in \{-1, 1\}$ such that

$$\mathbf{W}_{ij}(\lambda_r) \neq 0, \qquad \mathbf{W}_{ij}(\bar{\lambda}) = 0, \qquad \operatorname{sign}(\mathbf{W}_{ij}(\lambda_r)) \to s.$$

The active equality for the entering coordinate would then imply

$$\begin{aligned}
0 &= \lim_{r\to\infty}\left[\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda_r) + \mathbf{X}_{:,i}\mathbf{W}_{ij}(\lambda_r) - \mathbf{X}_{:,j}\right) + \lambda_r \operatorname{sign}(\mathbf{W}_{ij}(\lambda_r))\right] \\
&= \frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\bar{\lambda}) - \mathbf{X}_{:,j}\right) + \bar{\lambda}s.
\end{aligned}$$

Consequently,

$$\left|\frac{1}{n}\mathbf{X}_{:,i}^\top\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\bar{\lambda}) - \mathbf{X}_{:,j}\right)\right| = \bar{\lambda},$$

which is excluded by the non-transition hypothesis. Therefore every inactive entry remains zero on the same local regime: $\mathbf{W}_{ij}(\lambda) = 0$ for $(i,j) \notin \mathcal{A}$ and $\lambda \in I$. Since each inactive coordinate is the zero analytic function on I, $d\mathbf{W}_{ij}(\lambda)/d\lambda = 0$ for $(i,j) \notin \mathcal{A}$ and $\lambda \in I$. □

F Proof of Theorem 3.4

Proof. In this proof, all minima over empty index sets are interpreted as $+\infty$. Any inactive candidate containing a zero denominator is omitted. Let $\mathcal{I}_{\text{acyc}}$ be the set of inactive pairs $(i,j) \notin \mathcal{A}$ whose insertion $i \to j$ preserves acyclicity. For convenience, the symbol $\lambda_{ij}^{\text{enter}}$ denotes the quantity

$$\lambda_{ij}^{\text{enter}} = \max\left\{\frac{g_{ij} - \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}}, \frac{\lambda_0\dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}}\right\}.$$

On the current smooth branch the active set, the signs of the active entries, and the local acyclic regime are fixed. Hence, for every λ before the first transition point, we have

$$\mathbf{W}_{ij}(\lambda) = \mathbf{W}_{ij}(\lambda_0) + (\lambda - \lambda_0)\dot{\mathbf{W}}_{ij}, \qquad (i,j) \in \mathcal{A},$$

and $\mathbf{W}_{ij}(\lambda) = 0$ and $\dot{\mathbf{W}}_{ij} = 0$ for $(i,j) \notin \mathcal{A}$. For each j, the matrix-flow equation on the active coordinates is

$$\left(\frac{1}{n}\mathbf{X}_{:,\mathcal{A}_j}^\top\mathbf{X}_{:,\mathcal{A}_j}\right)\dot{\mathbf{W}}_{\mathcal{A}_j,j} = -\operatorname{sign}\left(\mathbf{W}_{\mathcal{A}_j,j}(\lambda_0)\right),$$

that is, entrywise,

$$\sum_{k\in\mathcal{A}_j}\frac{1}{n}\mathbf{X}_{:,i}^\top\mathbf{X}_{:,k}\dot{\mathbf{W}}_{kj} + \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda_0)\right) = 0, \qquad i \in \mathcal{A}_j.$$

For $i \in \mathcal{A}_j$, stationarity at λ_0 gives

$$\sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda_0) - \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda_0 \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda_0)\right) = 0.$$

For any λ such that the active sign of $\mathbf{W}_{ij}(\lambda)$ is still equal to the active sign of $\mathbf{W}_{ij}(\lambda_0)$, the active stationarity residual satisfies the exact identity

$$\begin{aligned}
&\sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda) - \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda)\right) \\
&= \sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \left[\mathbf{W}_{kj}(\lambda_0) + (\lambda - \lambda_0) \dot{\mathbf{W}}_{kj}\right] - \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda_0)\right) \\
&= \sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda_0) - \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda_0 \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda_0)\right) + (\lambda - \lambda_0) \left[\sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \dot{\mathbf{W}}_{kj} + \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda_0)\right)\right] \\
&= 0 + (\lambda - \lambda_0) \cdot 0 = 0.
\end{aligned}$$

Thus the active stationarity equations remain identically satisfied until an active sign changes. For $(i, j) \in \mathcal{A}$, active sign preservation is exactly the strict condition

$$\mathbf{W}_{ij}(\lambda_0) \left[\mathbf{W}_{ij}(\lambda_0) + (\lambda - \lambda_0) \dot{\mathbf{W}}_{ij}\right] > 0.$$

The boundary of this condition is $\mathbf{W}_{ij}(\lambda_0) + (\lambda - \lambda_0)\dot{\mathbf{W}}_{ij} = 0$. If $\dot{\mathbf{W}}_{ij} = 0$, then $\mathbf{W}_{ij}(\lambda) = \mathbf{W}_{ij}(\lambda_0) \neq 0$ for all λ on the current affine branch. If $\dot{\mathbf{W}}_{ij} \neq 0$, then the unique zero of the affine function $\lambda \mapsto \mathbf{W}_{ij}(\lambda)$ is $\lambda = \lambda_0 - \dfrac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}}$. For every active index with $\lambda_0 < \lambda_0 - \dfrac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}} \leq \lambda_{\max}$, the active sign is unchanged on $\left[\lambda_0, \lambda_0 - \dfrac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}}\right)$, and changes or becomes undefined at the endpoint. Taking the minimum over all such active indices gives

$$\lambda_{\text{drop}} = \min_{\substack{(i,j) \in \mathcal{A},\, \dot{\mathbf{W}}_{ij} \neq 0 \\ \lambda_0 < \lambda_0 - \mathbf{W}_{ij}(\lambda_0)/\dot{\mathbf{W}}_{ij} \leq \lambda_{\max}}} \left\{\lambda_0 - \frac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}}\right\}.$$

Consequently, for every λ satisfying $\lambda_0 < \lambda < \lambda_{\text{drop}}$, all active signs remain fixed, and for $\lambda = \lambda_{\text{drop}}$ at least one active coefficient reaches zero whenever $\lambda_{\text{drop}} < +\infty$.

Now take an inactive pair $(i, j) \notin \mathcal{A}$ whose insertion $i \to j$ preserves acyclicity. The inactive subgradient condition on the current regime is

$$\left|\frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right)\right| \leq \lambda.$$

Since the current branch is smooth and the current point is not already an inactive-entry transition, the inequality is strict at λ_0:

$$\left|\frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda_0) - \mathbf{X}_{:,j}\right)\right| < \lambda_0,$$

that is, $-\lambda_0 < g_{ij} < \lambda_0$. Along the current affine branch, the inactive correlation is exactly

$$\begin{aligned}
\frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j}\right) &= \frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \left[\mathbf{W}_{kj}(\lambda_0) + (\lambda - \lambda_0) \dot{\mathbf{W}}_{kj}\right] - \mathbf{X}_{:,j}\right) \\
&= \frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda_0) - \mathbf{X}_{:,j}\right) + (\lambda - \lambda_0) \frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \dot{\mathbf{W}}_{kj}\right) \\
&= g_{ij} + (\lambda - \lambda_0) \dot{g}_{ij}.
\end{aligned}$$

Hence the inactive subgradient condition is the pair of linear inequalities $-\lambda \le g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij} \le \lambda$, namely

$$g_{ij} + \lambda\dot{g}_{ij} - \lambda_0\dot{g}_{ij} - \lambda \le 0, \qquad -g_{ij} - \lambda\dot{g}_{ij} + \lambda_0\dot{g}_{ij} - \lambda \le 0,$$

or, written with the coefficients of λ separated,

$$g_{ij} - \lambda_0\dot{g}_{ij} - \lambda(1 - \dot{g}_{ij}) \le 0, \qquad -g_{ij} + \lambda_0\dot{g}_{ij} - \lambda(1 + \dot{g}_{ij}) \le 0.$$

The two boundary equations are therefore

$$g_{ij} - \lambda_0\dot{g}_{ij} - \lambda(1 - \dot{g}_{ij}) = 0, \qquad -g_{ij} + \lambda_0\dot{g}_{ij} - \lambda(1 + \dot{g}_{ij}) = 0.$$

If $1 - \dot{g}_{ij} \neq 0$, the first boundary equation gives $\lambda = \dfrac{g_{ij} - \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}}$. If $1 + \dot{g}_{ij} \neq 0$, the second boundary equation gives $\lambda = \dfrac{\lambda_0\dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}}$. The forward roots are identified as follows. Since $-\lambda_0 < g_{ij} < \lambda_0$, both numerators $g_{ij} - \lambda_0$ and $-g_{ij} - \lambda_0$ are strictly negative: $g_{ij} - \lambda_0 < 0$ and $-g_{ij} - \lambda_0 < 0$. For the first boundary root,

$$\begin{aligned}
\frac{g_{ij} - \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}} - \lambda_0 &= \frac{g_{ij} - \lambda_0\dot{g}_{ij} - \lambda_0(1 - \dot{g}_{ij})}{1 - \dot{g}_{ij}} \\
&= \frac{g_{ij} - \lambda_0\dot{g}_{ij} - \lambda_0 + \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}} \\
&= \frac{g_{ij} - \lambda_0}{1 - \dot{g}_{ij}}.
\end{aligned}$$

Thus, when $1 - \dot{g}_{ij} > 0$, this difference is strictly negative; when $1 - \dot{g}_{ij} < 0$, this difference is strictly positive. Therefore the first boundary root lies strictly above λ_0 exactly in the case $\dot{g}_{ij} > 1$. For the second boundary root,

$$\begin{aligned}
\frac{\lambda_0\dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}} - \lambda_0 &= \frac{\lambda_0\dot{g}_{ij} - g_{ij} - \lambda_0(1 + \dot{g}_{ij})}{1 + \dot{g}_{ij}} \\
&= \frac{\lambda_0\dot{g}_{ij} - g_{ij} - \lambda_0 - \lambda_0\dot{g}_{ij}}{1 + \dot{g}_{ij}} \\
&= \frac{-g_{ij} - \lambda_0}{1 + \dot{g}_{ij}}.
\end{aligned}$$

Thus, when $1 + \dot{g}_{ij} > 0$, this difference is strictly negative; when $1 + \dot{g}_{ij} < 0$, this difference is strictly positive. Therefore the second boundary root lies strictly above λ_0 exactly in the case $\dot{g}_{ij} < -1$. The two cases $\dot{g}_{ij} > 1$ and $\dot{g}_{ij} < -1$ are disjoint. If $-1 \le \dot{g}_{ij} \le 1$, no inactive boundary can be reached in the increasing-λ direction. Indeed, for every $\lambda \ge \lambda_0$,

$$\begin{aligned}
|g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij}| &\le |g_{ij}| + (\lambda - \lambda_0)|\dot{g}_{ij}| \\
&\le |g_{ij}| + (\lambda - \lambda_0) \\
&< \lambda_0 + (\lambda - \lambda_0) = \lambda.
\end{aligned}$$

If $\dot{g}_{ij} > 1$, then the upper inactive boundary is the only forward boundary. For every $\lambda_0 \le \lambda < \dfrac{g_{ij} - \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}}$, the upper residual satisfies

$$\begin{aligned}
g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij} - \lambda &= g_{ij} - \lambda_0\dot{g}_{ij} - \lambda(1 - \dot{g}_{ij}) \\
&< g_{ij} - \lambda_0\dot{g}_{ij} - \frac{g_{ij} - \lambda_0\dot{g}_{ij}}{1 - \dot{g}_{ij}}(1 - \dot{g}_{ij}) = 0,
\end{aligned}$$

while the lower residual is strictly positive as

$$\begin{aligned}
g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij} + \lambda &= g_{ij} - \lambda_0\dot{g}_{ij} + \lambda(1 + \dot{g}_{ij}) \\
&\ge g_{ij} - \lambda_0\dot{g}_{ij} + \lambda_0(1 + \dot{g}_{ij}) \\
&= g_{ij} + \lambda_0 > 0.
\end{aligned}$$

Thus, in the case $\dot{g}_{ij} > 1$, the first inactive equality reached in the increasing-λ direction is $\lambda = \dfrac{g_{ij} - \lambda_0 \dot{g}_{ij}}{1 - \dot{g}_{ij}}$. If $\dot{g}_{ij} < -1$, then the lower inactive boundary is the only forward boundary. For every $\lambda_0 \le \lambda < \dfrac{\lambda_0 \dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}}$, the lower residual satisfies

$$\begin{aligned} -g_{ij} - (\lambda - \lambda_0)\dot{g}_{ij} - \lambda &= -g_{ij} + \lambda_0 \dot{g}_{ij} - \lambda(1 + \dot{g}_{ij}) \\ &< -g_{ij} + \lambda_0 \dot{g}_{ij} - \frac{\lambda_0 \dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}}(1 + \dot{g}_{ij}) = 0, \end{aligned}$$

while the upper residual is strictly negative as

$$\begin{aligned} g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij} - \lambda &= g_{ij} - \lambda_0 \dot{g}_{ij} + \lambda(\dot{g}_{ij} - 1) \\ &\le g_{ij} - \lambda_0 \dot{g}_{ij} + \lambda_0(\dot{g}_{ij} - 1) \\ &= g_{ij} - \lambda_0 < 0. \end{aligned}$$

Thus, in the case $\dot{g}_{ij} < -1$, the first inactive equality reached in the increasing-λ direction is $\lambda = \dfrac{\lambda_0 \dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}}$. Combining the three cases $\dot{g}_{ij} > 1$, $-1 \le \dot{g}_{ij} \le 1$, and $\dot{g}_{ij} < -1$, and then retaining only candidate values in $(\lambda_0, \lambda_{\max}]$, the forward inactive-entry value for this inactive pair is precisely represented by

$$\lambda_{ij}^{\text{enter}} = \max \left\{ \frac{g_{ij} - \lambda_0 \dot{g}_{ij}}{1 - \dot{g}_{ij}}, \frac{\lambda_0 \dot{g}_{ij} - g_{ij}}{1 + \dot{g}_{ij}} \right\},$$

with zero-denominator candidates omitted. Taking the minimum over inactive pairs whose insertion preserves acyclicity gives

$$\lambda_{\text{enter}} = \min_{\substack{(i,j) \in \mathcal{I}_{\text{acyc}} \\ \lambda_0 < \lambda_{ij}^{\text{enter}} \le \lambda_{\max}}} \left\{ \lambda_{ij}^{\text{enter}} \right\}.$$

It remains to verify that the exact DAG constraint creates no additional scalar boundary while the support is unchanged. Since the current support is acyclic, by Appendix C there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{P}^\top \big(\mathbf{W}(\lambda) \odot \mathbf{W}(\lambda)\big)\mathbf{P}$ is strictly triangular for every λ before the first support change. Consequently,

$$\left[\mathbf{P}^\top \big(\mathbf{W}(\lambda) \odot \mathbf{W}(\lambda)\big)\mathbf{P}\right]^d = \mathbf{0},$$

and, for every integer $m \ge 1$,

$$\operatorname{tr}\left(\left[\mathbf{P}^\top \big(\mathbf{W}(\lambda) \odot \mathbf{W}(\lambda)\big)\mathbf{P}\right]^m\right) = 0.$$

Using similarity invariance of trace and the matrix exponential series, $h(\mathbf{W}(\lambda)) = 0$ holds identically on the unchanged acyclic support. An inactive pair whose insertion creates a directed cycle is not a feasible entering direction on the current acyclic regime, because the current active graph already contains a directed path $j \rightsquigarrow i$, and adding $i \to j$ produces the directed cycle $j \rightsquigarrow i \to j$. Such a pair can become eligible only after the active support changes.

Now let $\lambda_0 < \lambda < \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}$. The inequality $\lambda < \lambda_{\text{drop}}$ gives, for every $(i,j) \in \mathcal{A}$, $\mathbf{W}_{ij}(\lambda_0)\mathbf{W}_{ij}(\lambda) > 0$, and hence the active signs are unchanged. The active-stationarity identity above gives, for every $i \in \mathcal{A}_j$,

$$\sum_{k \in \mathcal{A}_j} \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda) - \frac{1}{n} \mathbf{X}_{:,i}^\top \mathbf{X}_{:,j} + \lambda \operatorname{sign}\left(\mathbf{W}_{ij}(\lambda)\right) = 0.$$

The inequality $\lambda < \lambda_{\text{enter}}$ gives, for every inactive pair $(i,j) \notin \mathcal{A}$ whose insertion preserves acyclicity, $|g_{ij} + (\lambda - \lambda_0)\dot{g}_{ij}| < \lambda$, and hence

$$\left| \frac{1}{n} \mathbf{X}_{:,i}^\top \left(\sum_{k \in \mathcal{A}_j} \mathbf{X}_{:,k} \mathbf{W}_{kj}(\lambda) - \mathbf{X}_{:,j} \right) \right| < \lambda.$$

Inactive pairs whose insertion creates a directed cycle are not feasible entering directions on the current acyclic regime, and the trace-exponential calculation above gives $h(\mathbf{W}(\lambda)) = 0$. Therefore, for every

$$\lambda \in [\lambda_0, \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}),$$

the same active set, the same active signs, and the same acyclic regime satisfy all stationary conditions of the current smooth branch. Hence no transition occurs before the displayed minimum. Finally, set

$$\lambda_{\text{next}} = \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}.$$

If $\lambda_{\text{next}} = \lambda_{\max}$, the current segment reaches the endpoint of the prescribed parameter interval before any transition. Suppose $\lambda_{\text{next}} < \lambda_{\max}$. If $\lambda_{\text{next}} = \lambda_{\text{drop}}$, then there exists $(i, j) \in \mathcal{A}$ such that $\lambda_{\text{next}} = \lambda_0 - \dfrac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}}$, and therefore

$$\begin{aligned}
\mathbf{W}_{ij}(\lambda_{\text{next}}) &= \mathbf{W}_{ij}(\lambda_0) + (\lambda_{\text{next}} - \lambda_0)\dot{\mathbf{W}}_{ij} \\
&= \mathbf{W}_{ij}(\lambda_0) + \left[\lambda_0 - \frac{\mathbf{W}_{ij}(\lambda_0)}{\dot{\mathbf{W}}_{ij}} - \lambda_0\right]\dot{\mathbf{W}}_{ij} \\
&= \mathbf{W}_{ij}(\lambda_0) - \mathbf{W}_{ij}(\lambda_0) = 0.
\end{aligned}$$

Thus the active sign representation fails at λ_{next}. If $\lambda_{\text{next}} = \lambda_{\text{enter}}$, then there exists an inactive pair $(i, j) \notin \mathcal{A}$ whose insertion preserves acyclicity and such that $\lambda_{\text{next}} = \lambda_{ij}^{\text{enter}}$. The preceding root calculation gives $|g_{ij} + (\lambda_{\text{next}} - \lambda_0)\dot{g}_{ij}| = \lambda_{\text{next}}$. Substituting the definitions of g_{ij}, $\dot{g}_{ij}$, and $\mathbf{W}_{kj}(\lambda_{\text{next}}) = \mathbf{W}_{kj}(\lambda_0) + (\lambda_{\text{next}} - \lambda_0)\dot{\mathbf{W}}_{kj}$, we obtain

$$\left|\frac{1}{n}\mathbf{X}_{:,i}^{\top}\left(\sum_{k\in\mathcal{A}_j}\mathbf{X}_{:,k}\mathbf{W}_{kj}(\lambda_{\text{next}}) - \mathbf{X}_{:,j}\right)\right| = \lambda_{\text{next}}.$$

Thus the inactive subgradient inequality is no longer strict at λ_{next}, and the inactive edge $i \to j$ reaches an acyclically admissible entry boundary. If both active-drop and inactive-entry events occur at the same value, or if multiple indices attain the same minimum, the transition is simultaneous. Therefore the nearest transition point in the increasing-λ direction is

$$\lambda_{\text{next}} = \min\{\lambda_{\max}, \lambda_{\text{drop}}, \lambda_{\text{enter}}\}.$$

□

G Proof of Theorem 4.1

Proof. We work on the effective domain of the GOLEM objective,

$$\mathcal{D} := \left\{\mathbf{W} \in \mathbb{R}^{d\times d} : \det(\mathbf{I} - \mathbf{W}) \neq 0,\ \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F > 0\right\}.$$

Throughout the proof, $\|\cdot\|_2$ denotes the spectral norm, and $\|\cdot\|_*$ the nuclear norm. We suppose that $\|\mathbf{X}\|_F > 0$, so that the logarithmic loss is finite at $\mathbf{W} = \mathbf{0}$. For fixed $\lambda_2 \geq 0$, write

$$F_{\lambda_1}(\mathbf{W}) := \frac{d}{2}\log\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I} - \mathbf{W})| + \lambda_1\|\mathbf{W}\|_1 + \lambda_2 h(\mathbf{W}), \quad \mathbf{W} \in \mathcal{D}.$$

Since $\mathbf{W} \odot \mathbf{W}$ is entrywise nonnegative, the matrix-exponential trace expansion (5) yields the nonnegativity identity

$$\begin{aligned}
h(\mathbf{W}) &= \sum_{k=1}^{\infty}\frac{1}{k!}\operatorname{tr}\big((\mathbf{W}\odot\mathbf{W})^k\big) \\
&= \sum_{k=1}^{\infty}\frac{1}{k!}\sum_{i_1,\ldots,i_k=1}^{d}(\mathbf{W}_{i_1i_2})^2(\mathbf{W}_{i_2i_3})^2\cdots(\mathbf{W}_{i_{k-1}i_k})^2(\mathbf{W}_{i_ki_1})^2 \geq 0.
\end{aligned} \tag{9}$$

First observe that $\det(\mathbf{I}) = 1$, $\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F = \|\mathbf{X}\|_F$, and $h(\mathbf{0}) = 0$, and therefore

$$F_{\lambda_1}(\mathbf{0}) = \frac{d}{2}\log\|\mathbf{X}\|_F^2.$$

For every $\mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}$, the denominator $\|\mathbf{W}\|_1$ is strictly positive. Hence the exact objective displacement from the origin is

$$\begin{aligned} F_{\lambda_1}(\mathbf{W}) - F_{\lambda_1}(\mathbf{0}) &= \frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \frac{d}{2}\log\|\mathbf{X}\|_F^2 - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_1\|\mathbf{W}\|_1 + \lambda_2 h(\mathbf{W}) \\ &= \lambda_1\|\mathbf{W}\|_1 + \frac{d}{2}\left[\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \log\|\mathbf{X}\|_F^2\right] - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_2 h(\mathbf{W}) \\ &= \lambda_1\|\mathbf{W}\|_1 - \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_2 h(\mathbf{W}) \\ &= \|\mathbf{W}\|_1\left[\lambda_1 - \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right]. \end{aligned} \tag{10}$$

We now prove the global-minimizer claim. If $\lambda_{\max} = +\infty$, then no finite λ_1 satisfies $\lambda_1 \geq \lambda_{\max}$, so the statement is vacuous. Assume $\lambda_{\max} < +\infty$. By the definition of $\lambda_{\max}$, for every $\mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}$,

$$\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1} \leq \lambda_{\max}. \tag{11}$$

Combining (10) and (11), for every admissible nonzero $\mathbf{W}$,

$$\begin{aligned} F_{\lambda_1}(\mathbf{W}) - F_{\lambda_1}(\mathbf{0}) &= \|\mathbf{W}\|_1\left[\lambda_1 - \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right] \\ &\geq \|\mathbf{W}\|_1(\lambda_1 - \lambda_{\max}). \end{aligned}$$

Thus, if $\lambda_1 \geq \lambda_{\max}$, then $F_{\lambda_1}(\mathbf{W}) \geq F_{\lambda_1}(\mathbf{0}), \quad \forall\, \mathbf{W} \in \mathcal{D}$, and $\mathbf{W} = \mathbf{0}$ is a global minimizer.

We next prove uniqueness when $\lambda_1 > \lambda_{\max}$. In this regime, for every $\mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}$,

$$\begin{aligned} F_{\lambda_1}(\mathbf{W}) - F_{\lambda_1}(\mathbf{0}) &\geq \|\mathbf{W}\|_1(\lambda_1 - \lambda_{\max}) \\ &> 0, \end{aligned}$$

because $\|\mathbf{W}\|_1 > 0$ for $\mathbf{W} \neq \mathbf{0}$. Hence no nonzero admissible matrix can attain the objective value at the origin, and $\mathbf{W} = \mathbf{0}$ is the unique global minimizer.

Moreover, the threshold is exact among nonnegative λ_1's. To see this, the factorization (10) gives

$$\begin{aligned} &\{\lambda_1 \in [0,\infty) : F_{\lambda_1}(\mathbf{W}) \geq F_{\lambda_1}(\mathbf{0}),\ \forall\, \mathbf{W} \in \mathcal{D}\} \\ &= \left\{\lambda_1 \in [0,\infty) : \|\mathbf{W}\|_1\left[\lambda_1 - \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right] \geq 0,\ \forall\, \mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}\right\} \\ &= \left\{\lambda_1 \in [0,\infty) : \lambda_1 \geq \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1},\ \forall\, \mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}\right\} \\ &= [0,\infty) \cap \bigcap_{\mathbf{W} \in \mathcal{D}\setminus\{\mathbf{0}\}} \left[\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}, \infty\right) \\ &= \left[\max\left\{0, \sup_{\substack{\mathbf{W}\neq\mathbf{0},\ \det(\mathbf{I}-\mathbf{W})\neq 0,\\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F > 0}} \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right\}, \infty\right) = [\lambda_{\max}, \infty). \end{aligned}$$

Equivalently, if $0 \leq \lambda_1 < \lambda_{\max} < +\infty$, then $\lambda_{\max} > 0$, and by the defining property of the supremum, for $\eta = (\lambda_{\max} - \lambda_1)/2 > 0$ there exists $\mathbf{W}_\eta \in \mathcal{D} \setminus \{\mathbf{0}\}$ such that

$$\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}_\eta\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W}_\eta)| - \lambda_2 h(\mathbf{W}_\eta)}{\|\mathbf{W}_\eta\|_1} > \lambda_{\max} - \eta = \lambda_1 + \eta,$$

and substituting this into (10) gives $F_{\lambda_1}(\mathbf{W}_\eta) - F_{\lambda_1}(\mathbf{0}) < -\eta\|\mathbf{W}_\eta\|_1 < 0$. Thus the threshold is not merely sufficient; it is the exact nonnegative regularization threshold for domination of all nonzero admissible matrices.

It remains to prove the finiteness characterization. First suppose that $\mathbf{X}^\top\mathbf{X} \succ 0$. For every $\mathbf{W} \in \mathcal{D}$, the matrix $\mathbf{I} - \mathbf{W}$ is invertible, and for every nonzero $v \in \mathbb{R}^d$,

$$v^\top(\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W})v = \{(\mathbf{I}-\mathbf{W})v\}^\top\mathbf{X}^\top\mathbf{X}\{(\mathbf{I}-\mathbf{W})v\} > 0,$$

because $(\mathbf{I}-\mathbf{W})v \neq 0$. Hence $(\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W}) \succ 0$. Moreover,

$$\begin{aligned}\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 &= \|\mathbf{X}(\mathbf{I}-\mathbf{W})\|_F^2 = \operatorname{tr}\big((\mathbf{X}(\mathbf{I}-\mathbf{W}))^\top\mathbf{X}(\mathbf{I}-\mathbf{W})\big)\\ &= \operatorname{tr}\big((\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W})\big).\end{aligned} \tag{12}$$

Let $\mu_1, \dots, \mu_d > 0$ be the eigenvalues of $(\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W})$. Then the arithmetic-geometric mean inequality and determinant multiplicativity give

$$\begin{aligned}\operatorname{tr}\big((\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W})\big) &= \sum_{j=1}^d \mu_j \geq d\left(\prod_{j=1}^d \mu_j\right)^{1/d}\\ &= d\det\big((\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W})\big)^{1/d}\\ &= d\left[\det((\mathbf{I}-\mathbf{W})^\top)\det(\mathbf{X}^\top\mathbf{X})\det(\mathbf{I}-\mathbf{W})\right]^{1/d}\\ &= d\left[\det(\mathbf{I}-\mathbf{W})^2\det(\mathbf{X}^\top\mathbf{X})\right]^{1/d}\\ &= d|\det(\mathbf{I}-\mathbf{W})|^{2/d}\det(\mathbf{X}^\top\mathbf{X})^{1/d}.\end{aligned} \tag{13}$$

Using $\lambda_2 \geq 0$, (9), (12), and (13), we get the global numerator estimate as

$$\begin{aligned}&\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})\\ &\leq \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\operatorname{tr}((\mathbf{I}-\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-\mathbf{W}))} + \log|\det(\mathbf{I}-\mathbf{W})|\\ &\leq \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d|\det(\mathbf{I}-\mathbf{W})|^{2/d}\det(\mathbf{X}^\top\mathbf{X})^{1/d}} + \log|\det(\mathbf{I}-\mathbf{W})|\\ &= \frac{d}{2}\log\|\mathbf{X}\|_F^2 - \frac{d}{2}\log d - \frac{d}{2}\cdot\frac{2}{d}\log|\det(\mathbf{I}-\mathbf{W})| - \frac{d}{2}\cdot\frac{1}{d}\log\det(\mathbf{X}^\top\mathbf{X}) + \log|\det(\mathbf{I}-\mathbf{W})|\\ &= \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d} - \frac{1}{2}\log\det(\mathbf{X}^\top\mathbf{X}).\end{aligned} \tag{14}$$

The global numerator bound does not control division by $\|\mathbf{W}\|_1$ near $\mathbf{W} = \mathbf{0}$. We therefore derive a local quotient estimate. Fix $\mathbf{W} \in \mathcal{D} \setminus \{\mathbf{0}\}$ with $0 < \|\mathbf{W}\|_1 \leq 1/2$. For every $t \in [0,1]$, finite-dimensional norm comparison, spectral-radius domination, and the Neumann series give

$$\begin{aligned}&\rho(t\mathbf{W}) \leq \|t\mathbf{W}\|_2 = t\|\mathbf{W}\|_2 \leq t\|\mathbf{W}\|_F = t\left(\sum_{i,j=1}^d W_{ij}^2\right)^{1/2} \leq t\sum_{i,j=1}^d |W_{ij}| = t\|\mathbf{W}\|_1 \leq \frac{1}{2},\\ &(\mathbf{I}-t\mathbf{W})^{-1} = \sum_{k=0}^\infty (t\mathbf{W})^k, \qquad \|(\mathbf{I}-t\mathbf{W})^{-1}\|_2 \leq \sum_{k=0}^\infty t^k\|\mathbf{W}\|_2^k = \frac{1}{1-t\|\mathbf{W}\|_2} \leq \frac{1}{1-\|\mathbf{W}\|_1} \leq 2,\\ &\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F = \|\mathbf{X}-t\mathbf{X}\mathbf{W}\|_F \geq \|\mathbf{X}\|_F - t\|\mathbf{X}\mathbf{W}\|_F \geq \|\mathbf{X}\|_F - t\|\mathbf{X}\|_F\|\mathbf{W}\|_2\\ &\qquad = \|\mathbf{X}\|_F(1-t\|\mathbf{W}\|_2) \geq \|\mathbf{X}\|_F(1-\|\mathbf{W}\|_1) \geq \frac{1}{2}\|\mathbf{X}\|_F.\end{aligned} \tag{15}$$

In particular, $t \mapsto \mathbf{I} - t\mathbf{W}$ remains inside $GL_d(\mathbb{R})$ [Horn and Johnson, 2012], and $\det(\mathbf{I}-t\mathbf{W})$ has constant positive sign because $\det(\mathbf{I}) = 1$. Hence the scalar map

$$t \mapsto \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2} + \log|\det(\mathbf{I}-t\mathbf{W})|$$

is continuously differentiable on $[0,1]$. Applying the fundamental theorem of calculus, differentiating the Frobenius norm term, and using Jacobi's formula for the determinant, we obtain

$$
\begin{aligned}
&\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}(\mathbf{I}-\mathbf{W})\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})| \\
&=\int_0^1 \frac{d}{dt}\left\{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}+\log|\det(\mathbf{I}-t\mathbf{W})|\right\}dt \\
&=\int_0^1\left[-\frac{d}{2}\frac{\frac{d}{dt}\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}+\frac{\frac{d}{dt}\det(\mathbf{I}-t\mathbf{W})}{\det(\mathbf{I}-t\mathbf{W})}\right]dt \\
&=\int_0^1\left[-\frac{d}{2}\frac{\frac{d}{dt}\operatorname{tr}\big((\mathbf{I}-t\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-t\mathbf{W})\big)}{\operatorname{tr}((\mathbf{I}-t\mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-t\mathbf{W}))}+\operatorname{tr}\big((\mathbf{I}-t\mathbf{W})^{-1}(-\mathbf{W})\big)\right]dt \\
&=\int_0^1\left[-\frac{d}{2}\frac{-2\operatorname{tr}\big(\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}-t\mathbf{W})\big)}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}-\operatorname{tr}\big((\mathbf{I}-t\mathbf{W})^{-1}\mathbf{W}\big)\right]dt \\
&=\int_0^1\left[d\frac{\langle\mathbf{X}(\mathbf{I}-t\mathbf{W}),\mathbf{X}\mathbf{W}\rangle_F}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}-\operatorname{tr}\big((\mathbf{I}-t\mathbf{W})^{-1}\mathbf{W}\big)\right]dt.
\end{aligned}
\tag{16}
$$

Now combine $\lambda_2 h(\mathbf{W})\geq 0$, (16), Cauchy-Schwarz, submultiplicativity, $\|\mathbf{X}\mathbf{W}\|_F\leq\|\mathbf{X}\|_F\|\mathbf{W}\|_2$, $|\operatorname{tr}(\mathbf{A}\mathbf{B})|\leq\|\mathbf{A}\|_2\|\mathbf{B}\|_*\leq d\|\mathbf{A}\|_2\|\mathbf{B}\|_2$, and (15). For $0<\|\mathbf{W}\|_1\leq 1/2$,

$$
\begin{aligned}
&\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|-\lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1} \\
&\leq\frac{1}{\|\mathbf{W}\|_1}\left|\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}(\mathbf{I}-\mathbf{W})\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|\right| \\
&\leq\frac{1}{\|\mathbf{W}\|_1}\int_0^1\left[d\frac{|\langle\mathbf{X}(\mathbf{I}-t\mathbf{W}),\mathbf{X}\mathbf{W}\rangle_F|}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}+\left|\operatorname{tr}\big((\mathbf{I}-t\mathbf{W})^{-1}\mathbf{W}\big)\right|\right]dt \\
&\leq\frac{1}{\|\mathbf{W}\|_1}\int_0^1\left[d\frac{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F\,\|\mathbf{X}\mathbf{W}\|_F}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F^2}+\|(\mathbf{I}-t\mathbf{W})^{-1}\|_2\,\|\mathbf{W}\|_*\right]dt \\
&\leq\frac{1}{\|\mathbf{W}\|_1}\int_0^1\left[d\frac{\|\mathbf{X}\|_F\|\mathbf{W}\|_2}{\|\mathbf{X}(\mathbf{I}-t\mathbf{W})\|_F}+d\|(\mathbf{I}-t\mathbf{W})^{-1}\|_2\|\mathbf{W}\|_2\right]dt \\
&\leq\frac{d}{\|\mathbf{W}\|_1}\int_0^1\left[\frac{\|\mathbf{X}\|_F\|\mathbf{W}\|_2}{\|\mathbf{X}\|_F(1-t\|\mathbf{W}\|_2)}+\frac{\|\mathbf{W}\|_2}{1-t\|\mathbf{W}\|_2}\right]dt \\
&=\frac{2d}{\|\mathbf{W}\|_1}\int_0^1\frac{\|\mathbf{W}\|_2}{1-t\|\mathbf{W}\|_2}\,dt=\frac{2d}{\|\mathbf{W}\|_1}\left[-\log(1-t\|\mathbf{W}\|_2)\right]_0^1 \\
&=\frac{2d}{\|\mathbf{W}\|_1}\{-\log(1-\|\mathbf{W}\|_2)\}=\frac{2d}{\|\mathbf{W}\|_1}\int_0^{\|\mathbf{W}\|_2}\frac{1}{1-s}\,ds \\
&\leq\frac{2d}{\|\mathbf{W}\|_1}\int_0^{\|\mathbf{W}\|_2}\frac{1}{1-\|\mathbf{W}\|_2}\,ds=\frac{2d}{\|\mathbf{W}\|_1}\frac{\|\mathbf{W}\|_2}{1-\|\mathbf{W}\|_2} \\
&\leq\frac{2d}{\|\mathbf{W}\|_1}\frac{\|\mathbf{W}\|_F}{1-\|\mathbf{W}\|_F} \\
&\leq\frac{2d}{\|\mathbf{W}\|_1}\frac{\|\mathbf{W}\|_1}{1-\|\mathbf{W}\|_1}\leq 4d.
\end{aligned}
\tag{17}
$$

For the complementary region $\|\mathbf{W}\|_1>1/2$, the global bound (14) gives

$$\begin{aligned}\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|-\lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1} &\le \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d}-\frac{1}{2}\log\det(\mathbf{X}^\top\mathbf{X})}{\|\mathbf{W}\|_1}\\ &\le \frac{\max\left\{0,\ \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d}-\frac{1}{2}\log\det(\mathbf{X}^\top\mathbf{X})\right\}}{\|\mathbf{W}\|_1}\\ &< 2\max\left\{0,\ \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d}-\frac{1}{2}\log\det(\mathbf{X}^\top\mathbf{X})\right\}.\end{aligned} \tag{18}$$

Combining (17) and (18), we obtain the finite supremum estimate

$$\begin{aligned}&\sup_{\substack{\mathbf{W}\neq\mathbf{0},\ \det(\mathbf{I}-\mathbf{W})\neq 0,\\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F>0}}\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|-\lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\\ &\le \max\left\{\sup_{\substack{\mathbf{W}\neq\mathbf{0},\ \det(\mathbf{I}-\mathbf{W})\neq 0,\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F>0\\ 0<\|\mathbf{W}\|_1\le 1/2}}\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|-\lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1},\right.\\ &\qquad\left.\sup_{\substack{\mathbf{W}\neq\mathbf{0},\ \det(\mathbf{I}-\mathbf{W})\neq 0,\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F>0\\ \|\mathbf{W}\|_1>1/2}}\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\log|\det(\mathbf{I}-\mathbf{W})|-\lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1}\right\}\\ &\le \max\left\{4d,\ 2\max\left\{0,\ \frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{d}-\frac{1}{2}\log\det(\mathbf{X}^\top\mathbf{X})\right\}\right\}<+\infty.\end{aligned} \tag{19}$$

Thus $\lambda_{\max}<+\infty$ whenever $\mathbf{X}^\top\mathbf{X}\succ 0$.

Conversely, suppose $\mathbf{X}^\top\mathbf{X}$ is singular. Since $\|\mathbf{X}\|_F>0$, set $r:=\operatorname{rank}(\mathbf{X})\in\{1,\dots,d-1\}$. Let $\mathbf{P}$ be the orthogonal projector onto the row space of $\mathbf{X}$. Then $\mathbf{P}^\top=\mathbf{P}$, $\mathbf{P}^2=\mathbf{P}$, $\operatorname{rank}(\mathbf{P})=r$, $\mathbf{X}\mathbf{P}=\mathbf{X}$, and $\mathbf{X}(\mathbf{I}-\mathbf{P})=\mathbf{0}$. For $\varepsilon\in(0,1)$, define $\mathbf{W}_\varepsilon:=(1-\varepsilon)\mathbf{P}$. Because $\mathbf{P}$ is an orthogonal projector, there exists an orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{P}=\mathbf{U}\begin{bmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}^\top,\qquad \mathbf{I}-\mathbf{W}_\varepsilon=\varepsilon\mathbf{P}+(\mathbf{I}-\mathbf{P})=\mathbf{U}\begin{bmatrix}\varepsilon\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{d-r}\end{bmatrix}\mathbf{U}^\top.$$

Therefore

$$\begin{aligned}\det(\mathbf{I}-\mathbf{W}_\varepsilon)&=\det(\mathbf{U})\det\begin{bmatrix}\varepsilon\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{d-r}\end{bmatrix}\det(\mathbf{U}^\top)=\varepsilon^r\neq 0,\\ \mathbf{X}-\mathbf{X}\mathbf{W}_\varepsilon&=\mathbf{X}(\mathbf{I}-\mathbf{W}_\varepsilon)=\mathbf{X}\{\varepsilon\mathbf{P}+(\mathbf{I}-\mathbf{P})\}=\varepsilon\mathbf{X}\mathbf{P}+\mathbf{X}(\mathbf{I}-\mathbf{P})=\varepsilon\mathbf{X},\\ \|\mathbf{X}-\mathbf{X}\mathbf{W}_\varepsilon\|_F^2&=\varepsilon^2\|\mathbf{X}\|_F^2>0.\end{aligned} \tag{20}$$

Hence $\mathbf{W}_\varepsilon\in\mathcal{D}\setminus\{\mathbf{0}\}$ for every $\varepsilon\in(0,1)$. Moreover,

$$\begin{aligned}\|\mathbf{W}_\varepsilon\|_1&=(1-\varepsilon)\|\mathbf{P}\|_1,\\ \|\mathbf{P}\|_1&=\sum_{i,j=1}^d|P_{ij}|\ge\sum_{i=1}^d|P_{ii}|=\sum_{i=1}^d P_{ii}=\operatorname{tr}(\mathbf{P})=r>0,\end{aligned} \tag{21}$$

where $P_{ii}\ge 0$ follows from $\mathbf{P}\succeq 0$. We next prove that the acyclicity penalty is uniformly bounded along this singular path. Since $\mathbf{W}_\varepsilon\odot\mathbf{W}_\varepsilon=(1-\varepsilon)^2(\mathbf{P}\odot\mathbf{P})$, the exponential series, trace (operator)

norm inequalities, submultiplicativity, and $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F$ yield

$$\begin{aligned} h(\mathbf{W}_\varepsilon) &= \operatorname{tr}\big(e^{\mathbf{W}_\varepsilon \odot \mathbf{W}_\varepsilon}\big) - d = \operatorname{tr}\Big(e^{(1-\varepsilon)^2(\mathbf{P}\odot\mathbf{P})}\Big) - d \\ &= \sum_{k=1}^{\infty} \frac{(1-\varepsilon)^{2k}}{k!} \operatorname{tr}\big((\mathbf{P}\odot\mathbf{P})^k\big) \\ &\le \sum_{k=1}^{\infty} \frac{1}{k!} \left|\operatorname{tr}\big((\mathbf{P}\odot\mathbf{P})^k\big)\right| \le \sum_{k=1}^{\infty} \frac{d}{k!} \left\|(\mathbf{P}\odot\mathbf{P})^k\right\|_2 \\ &\le d \sum_{k=1}^{\infty} \frac{\|\mathbf{P}\odot\mathbf{P}\|_2^k}{k!} \le d \sum_{k=1}^{\infty} \frac{\|\mathbf{P}\odot\mathbf{P}\|_F^k}{k!} \\ &= d\left(e^{\|\mathbf{P}\odot\mathbf{P}\|_F} - 1\right). \end{aligned} \tag{22}$$

Furthermore,

$$\begin{aligned} \|\mathbf{P}\odot\mathbf{P}\|_F = \left(\sum_{i,j=1}^{d} P_{ij}^4\right)^{1/2} &\le \sum_{i,j=1}^{d} P_{ij}^2 = \|\mathbf{P}\|_F^2 \\ &= \operatorname{tr}(\mathbf{P}^\top\mathbf{P}) = \operatorname{tr}(\mathbf{P}^2) = \operatorname{tr}(\mathbf{P}) = r. \end{aligned} \tag{23}$$

Combining (22) and (23), we get

$$0 \le h(\mathbf{W}_\varepsilon) \le d(e^r - 1), \qquad \forall\, \varepsilon \in (0,1). \tag{24}$$

Substituting (20), (21), and (24) into the threshold quotient gives

$$\begin{aligned} \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}_\varepsilon\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W}_\varepsilon)| - \lambda_2 h(\mathbf{W}_\varepsilon)}{\|\mathbf{W}_\varepsilon\|_1} &= \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\varepsilon^2\|\mathbf{X}\|_F^2} + \log(\varepsilon^r) - \lambda_2 h((1-\varepsilon)\mathbf{P})}{(1-\varepsilon)\|\mathbf{P}\|_1} \\ &= \frac{\frac{d}{2}\log(\varepsilon^{-2}) + r\log\varepsilon - \lambda_2 h((1-\varepsilon)\mathbf{P})}{(1-\varepsilon)\|\mathbf{P}\|_1} \\ &= \frac{d\log(\varepsilon^{-1}) - r\log(\varepsilon^{-1}) - \lambda_2 h((1-\varepsilon)\mathbf{P})}{(1-\varepsilon)\|\mathbf{P}\|_1} \\ &= \frac{(d-r)\log(\varepsilon^{-1}) - \lambda_2 h((1-\varepsilon)\mathbf{P})}{(1-\varepsilon)\|\mathbf{P}\|_1} \\ &\ge \frac{(d-r)\log(\varepsilon^{-1}) - \lambda_2 d(e^r-1)}{(1-\varepsilon)\|\mathbf{P}\|_1}. \end{aligned} \tag{25}$$

For sufficiently small ε, namely whenever $\log(\varepsilon^{-1}) \ge \dfrac{2\lambda_2 d(e^r-1)}{d-r}$, the numerator in the final line of (25) satisfies $(d-r)\log(\varepsilon^{-1}) - \lambda_2 d(e^r-1) \ge \dfrac{d-r}{2}\log(\varepsilon^{-1})$. Since $1-\varepsilon \le 1$, $\|\mathbf{P}\|_1 > 0$, and the numerator above is positive in this regime, (25) further gives

$$\begin{aligned} &\frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}_\varepsilon\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W}_\varepsilon)| - \lambda_2 h(\mathbf{W}_\varepsilon)}{\|\mathbf{W}_\varepsilon\|_1} \\ &\ge \frac{\frac{d-r}{2}\log(\varepsilon^{-1})}{(1-\varepsilon)\|\mathbf{P}\|_1} \\ &\ge \frac{d-r}{2\|\mathbf{P}\|_1}\log(\varepsilon^{-1}) \xrightarrow[\varepsilon\downarrow 0]{} +\infty. \end{aligned} \tag{26}$$

Therefore

$$\sup_{\substack{\mathbf{W}\neq\mathbf{0},\ \det(\mathbf{I}-\mathbf{W})\neq 0,\\ \|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F>0}} \frac{\frac{d}{2}\log\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \log|\det(\mathbf{I}-\mathbf{W})| - \lambda_2 h(\mathbf{W})}{\|\mathbf{W}\|_1} = +\infty,$$

and thus $\lambda_{\max} = +\infty$ whenever $\mathbf{X}^\top\mathbf{X}$ is singular. Combining the positive-definite direction (19) with the singular construction (26), we conclude that $\lambda_{\max} < +\infty \iff \mathbf{X}^\top\mathbf{X} \succ 0$. The proof is complete. □

H Proof of Theorem 4.2

Proof. We denote $\mathbf{W}^0 := \mathbf{W}(\lambda_1^0, \lambda_2^0)$ and $\mathcal{A} := \{(i,j) : [\mathbf{W}^0]_{ij} \neq 0\}$, and let $\mathbf{1}_{\{P\}}$ denote the indicator of the proposition P. In this proof, the assertion that $(\mathbf{W}^0, \lambda_1^0, \lambda_2^0)$ is not a transition point is used in the regular active-stratum sense: the ℓ_1-active signs are stable, the inactive ℓ_1-subgradient inequalities are strict, and the active linearization of the smooth GOLEM stationarity field is nonsingular. Thus we know

$$\left| -d\frac{\left[\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^0\|_F^2} + \left[(\mathbf{I} - \mathbf{W}^0)^{-\top}\right]_{ij} + 2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0 \odot \mathbf{W}^0}\right]_{jj} \right| < \lambda_1^0, \qquad (i,j) \notin \mathcal{A},$$

and

$$\det\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I} - \mathbf{W})| + \lambda_2^0 h(\mathbf{W})\right\}\Big|_{\mathbf{W}=\mathbf{W}^0}\right]_{(i,j),(k,l)\in\mathcal{A}} \neq 0,$$

with the determinant understood as 1 when $\mathcal{A} = \varnothing$. We begin by recording the complete differential structure of the smooth part of (3). Since $\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 = \operatorname{tr}\big((\mathbf{I} - \mathbf{W})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \mathbf{W})\big)$, for every perturbation $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0}\|\mathbf{X} - \mathbf{X}(\mathbf{W} + t\mathbf{\Delta})\|_F^2 &= \frac{d}{dt}\Big|_{t=0}\operatorname{tr}\big((\mathbf{I} - \mathbf{W} - t\mathbf{\Delta})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \mathbf{W} - t\mathbf{\Delta})\big)\\ &= -\operatorname{tr}\big(\mathbf{\Delta}^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \mathbf{W})\big) - \operatorname{tr}\big((\mathbf{I} - \mathbf{W})^\top\mathbf{X}^\top\mathbf{X}\mathbf{\Delta}\big)\\ &= -2\left\langle\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \mathbf{W}), \mathbf{\Delta}\right\rangle_F = -2\left\langle\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W}), \mathbf{\Delta}\right\rangle_F,\end{aligned}$$

and hence

$$\begin{aligned}&\frac{d}{dt}\Big|_{t=0}\frac{d}{2}\log\|\mathbf{X} - \mathbf{X}(\mathbf{W} + t\mathbf{\Delta})\|_F^2\\ &= \frac{d}{2}\frac{-2\left\langle\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W}), \mathbf{\Delta}\right\rangle_F}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} = \left\langle -d\frac{\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}, \mathbf{\Delta}\right\rangle_F.\end{aligned}$$

Similarly, since $\det(\mathbf{I} - \mathbf{W}^0) \neq 0$, the determinant has constant sign in a sufficiently small neighborhood of $\mathbf{W}^0$, and

$$\begin{aligned}&\frac{d}{dt}\Big|_{t=0}\{-\log|\det(\mathbf{I} - \mathbf{W} - t\mathbf{\Delta})|\}\\ &= -\frac{d}{dt}\Big|_{t=0}\log|\det(\mathbf{I} - \mathbf{W} - t\mathbf{\Delta})|\\ &= -\operatorname{tr}\big((\mathbf{I} - \mathbf{W})^{-1}(-\mathbf{\Delta})\big) = \operatorname{tr}\big((\mathbf{I} - \mathbf{W})^{-1}\mathbf{\Delta}\big) = \left\langle(\mathbf{I} - \mathbf{W})^{-\top}, \mathbf{\Delta}\right\rangle_F.\end{aligned}$$

The acyclicity penalty requires the full noncommutative exponential calculus. Temporarily writing $\mathbf{M} = \mathbf{W} \odot \mathbf{W}$ only inside the present calculation, we know

$$h(\mathbf{W}) = \operatorname{tr}(e^{\mathbf{W}\odot\mathbf{W}}) - d = \sum_{q=1}^{\infty}\frac{\operatorname{tr}((\mathbf{W}\odot\mathbf{W})^q)}{q!}.$$

Because $\mathbf{W} \odot \mathbf{W}$ is entrywise nonnegative, every diagonal entry of $(\mathbf{W} \odot \mathbf{W})^q$ is nonnegative. Moreover,

$$\begin{aligned}0 \le h(\mathbf{W}) = \sum_{q=1}^{\infty}\frac{\operatorname{tr}((\mathbf{W}\odot\mathbf{W})^q)}{q!} &\le \sum_{q=1}^{\infty}\frac{d\|\mathbf{W}\odot\mathbf{W}\|_2^q}{q!} \le d\sum_{q=1}^{\infty}\frac{\|\mathbf{W}\odot\mathbf{W}\|_F^q}{q!}\\ &\le d\sum_{q=1}^{\infty}\frac{\|\mathbf{W}\|_F^{2q}}{q!} = d\left(e^{\|\mathbf{W}\|_F^2} - 1\right).\end{aligned}$$

For any perturbation $\mathbf{N}$, the Fréchet derivative of the exponential follows from the series expansion

$$\begin{aligned}\lim_{t\to 0}\frac{e^{\mathbf{M}+t\mathbf{N}}-e^{\mathbf{M}}}{t} &= \lim_{t\to 0}\sum_{q=1}^{\infty}\frac{1}{q!}\frac{(\mathbf{M}+t\mathbf{N})^q-\mathbf{M}^q}{t}\\ &= \sum_{q=1}^{\infty}\frac{1}{q!}\sum_{r=0}^{q-1}\mathbf{M}^r\mathbf{N}\mathbf{M}^{q-1-r}\\ &= \int_0^1 e^{(1-s)\mathbf{M}}\mathbf{N}e^{s\mathbf{M}}\,ds,\end{aligned}$$

where the last identity is obtained by expanding both exponentials and using $\int_0^1 (1-s)^a s^b\,ds = \frac{a!\,b!}{(a+b+1)!}$ [Higham, 2008]. Thus

$$\begin{aligned}\left.\frac{d}{dt}\right|_{t=0} h(\mathbf{W}+t\boldsymbol{\Delta}) &= \operatorname{tr}\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\boldsymbol{\Delta})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds\right)\\ &= \int_0^1 \operatorname{tr}\left(e^{s(\mathbf{W}\odot\mathbf{W})}e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\boldsymbol{\Delta})\right)ds\\ &= \sum_{i,j=1}^{d} 2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}[\boldsymbol{\Delta}]_{ij} = \left\langle 2\mathbf{W}\odot(e^{\mathbf{W}\odot\mathbf{W}})^\top, \boldsymbol{\Delta}\right\rangle_F.\end{aligned}$$

Consequently, $\frac{\partial h(\mathbf{W})}{\partial[\mathbf{W}]_{ij}} = 2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}$. Combining the residual, log-determinant, and acyclicity parts gives the full smooth stationarity field

$$\begin{aligned}&\frac{\partial}{\partial[\mathbf{W}]_{ij}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_2 h(\mathbf{W})\right\}\\ &\qquad = -d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} + 2\lambda_2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}.\end{aligned}$$

The second derivative is equally explicit. Differentiating the residual contribution entrywise gives

$$\begin{aligned}&\frac{\partial}{\partial[\mathbf{W}]_{kl}}\left(-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right)\\ &= -d\frac{-\left[\mathbf{X}^\top\mathbf{X}\right]_{ik}\mathbf{1}_{\{j=l\}}\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}\left(-2\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{kl}\right)}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4}\\ &= d\frac{\left[\mathbf{X}^\top\mathbf{X}\right]_{ik}\mathbf{1}_{\{j=l\}}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - 2d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{kl}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4}.\end{aligned}$$

For the determinant contribution,

$$\begin{aligned}\frac{\partial}{\partial[\mathbf{W}]_{kl}}\left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij} &= \left[(\mathbf{I}-\mathbf{W})^{-\top}\left(\frac{\partial\mathbf{W}}{\partial[\mathbf{W}]_{kl}}\right)^\top(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}\\ &= \left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{il}\left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{kj}.\end{aligned}$$

For the acyclicity part, the matrix-exponential derivative gives

$$\begin{aligned}&\frac{\partial}{\partial[\mathbf{W}]_{kl}}\left(2\lambda_2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}\right)\\ &= 2\lambda_2\mathbf{1}_{\{i=k\}}\mathbf{1}_{\{j=l\}}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji} + 2\lambda_2[\mathbf{W}]_{ij}\left[\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2[\mathbf{W}]_{kl}\mathbf{e}_k\mathbf{e}_l^\top)e^{s(\mathbf{W}\odot\mathbf{W})}\,ds\right]_{ji}\\ &= 2\lambda_2\mathbf{1}_{\{i=k\}}\mathbf{1}_{\{j=l\}}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji} + 4\lambda_2[\mathbf{W}]_{ij}[\mathbf{W}]_{kl}\int_0^1\left[e^{(1-s)(\mathbf{W}\odot\mathbf{W})}\right]_{jk}\left[e^{s(\mathbf{W}\odot\mathbf{W})}\right]_{li}ds,\end{aligned}$$

where $\mathbf{e}_k$ is the k-th coordinate vector. Hence, for all $1 \le i, j, k, l \le d$,

$$\begin{aligned}
&\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}} \left\{ \frac{d}{2} \log \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I} - \mathbf{W})| + \lambda_2 h(\mathbf{W}) \right\} \\
&= d\frac{\left[\mathbf{X}^\top \mathbf{X}\right]_{ik} \mathbf{1}_{\{j=l\}}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} - 2d\frac{\left[\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W})\right]_{ij}\left[\mathbf{X}^\top(\mathbf{X} - \mathbf{X}\mathbf{W})\right]_{kl}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4} + \left[(\mathbf{I} - \mathbf{W})^{-\top}\right]_{il}\left[(\mathbf{I} - \mathbf{W})^{-\top}\right]_{kj} \\
&\quad + 4\lambda_2[\mathbf{W}]_{ij}[\mathbf{W}]_{kl}\int_0^1 \left[e^{(1-s)(\mathbf{W}\odot\mathbf{W})}\right]_{jk}\left[e^{s(\mathbf{W}\odot\mathbf{W})}\right]_{li} ds + 2\lambda_2 \mathbf{1}_{\{i=k\}}\mathbf{1}_{\{j=l\}}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}.
\end{aligned}$$

The preceding computations also yield the local analytic bounds needed to preserve the nonsmooth stratum. For $\|\mathbf{W} - \mathbf{W}^0\|_F \le r$,

$$\begin{aligned}
\left|\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}^0\|_F^2\right| &= \left|\|\mathbf{X}(\mathbf{I} - \mathbf{W}^0) - \mathbf{X}(\mathbf{W} - \mathbf{W}^0)\|_F^2 - \|\mathbf{X}(\mathbf{I} - \mathbf{W}^0)\|_F^2\right| \\
&\le 2\|\mathbf{X}(\mathbf{I} - \mathbf{W}^0)\|_F\|\mathbf{X}(\mathbf{W} - \mathbf{W}^0)\|_F + \|\mathbf{X}(\mathbf{W} - \mathbf{W}^0)\|_F^2 \\
&\le 2\|\mathbf{X} - \mathbf{X}\mathbf{W}^0\|_F\|\mathbf{X}\|_2\|\mathbf{W} - \mathbf{W}^0\|_F + \|\mathbf{X}\|_2^2\|\mathbf{W} - \mathbf{W}^0\|_F^2 \\
&\le \left(2\|\mathbf{X} - \mathbf{X}\mathbf{W}^0\|_F\|\mathbf{X}\|_2 + r\|\mathbf{X}\|_2^2\right)\|\mathbf{W} - \mathbf{W}^0\|_F.
\end{aligned}$$

Thus, for sufficiently small r, $\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 \ge \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{W}^0\|_F^2$. For the determinant part, whenever $\|(\mathbf{I} - \mathbf{W}^0)^{-1}\|_2\|\mathbf{W} - \mathbf{W}^0\|_F < 1$, one has

$$\begin{aligned}
\mathbf{I} - \mathbf{W} &= (\mathbf{I} - \mathbf{W}^0)\left[\mathbf{I} - (\mathbf{I} - \mathbf{W}^0)^{-1}(\mathbf{W} - \mathbf{W}^0)\right], \\
(\mathbf{I} - \mathbf{W})^{-1} &= \left[\sum_{q=0}^{\infty}\left\{(\mathbf{I} - \mathbf{W}^0)^{-1}(\mathbf{W} - \mathbf{W}^0)\right\}^q\right](\mathbf{I} - \mathbf{W}^0)^{-1},
\end{aligned}$$

and hence

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{W})^{-1}\|_2 &\le \sum_{q=0}^{\infty}\|(\mathbf{I} - \mathbf{W}^0)^{-1}\|_2^{q+1}\|\mathbf{W} - \mathbf{W}^0\|_F^q \\
&= \frac{\|(\mathbf{I} - \mathbf{W}^0)^{-1}\|_2}{1 - \|(\mathbf{I} - \mathbf{W}^0)^{-1}\|_2\|\mathbf{W} - \mathbf{W}^0\|_F},
\end{aligned}$$

while the determinant lower bound is

$$\begin{aligned}
|\det(\mathbf{I} - \mathbf{W})| &= |\det(\mathbf{I} - \mathbf{W}^0)|\left|\det\left[\mathbf{I} - (\mathbf{I} - \mathbf{W}^0)^{-1}(\mathbf{W} - \mathbf{W}^0)\right]\right| \\
&\ge |\det(\mathbf{I} - \mathbf{W}^0)|\prod_{q=1}^{d}\left(1 - \|(\mathbf{I} - \mathbf{W}^0)^{-1}(\mathbf{W} - \mathbf{W}^0)\|_2\right) \\
&\ge |\det(\mathbf{I} - \mathbf{W}^0)|\left(1 - \|(\mathbf{I} - \mathbf{W}^0)^{-1}\|_2\|\mathbf{W} - \mathbf{W}^0\|_F\right)^d.
\end{aligned}$$

Thus, for smaller r if necessary, $|\det(\mathbf{I} - \mathbf{W})| \ge \frac{1}{2}|\det(\mathbf{I} - \mathbf{W}^0)|$. The acyclicity contribution is controlled at the same level. Since

$$\begin{aligned}
\|(\mathbf{W}\odot\mathbf{W}) - (\mathbf{W}^0\odot\mathbf{W}^0)\|_F &= \|(\mathbf{W} - \mathbf{W}^0)\odot\mathbf{W} + \mathbf{W}^0\odot(\mathbf{W} - \mathbf{W}^0)\|_F \\
&\le \|\mathbf{W} - \mathbf{W}^0\|_F\|\mathbf{W}\|_F + \|\mathbf{W}^0\|_F\|\mathbf{W} - \mathbf{W}^0\|_F \\
&\le (2\|\mathbf{W}^0\|_F + r)\|\mathbf{W} - \mathbf{W}^0\|_F,
\end{aligned}$$

and

$$\begin{aligned}
\|e^{\mathbf{W}\odot\mathbf{W}} - e^{\mathbf{W}^0\odot\mathbf{W}^0}\|_F &= \left\|\int_0^1 \mathrm{D}e^{\mathbf{W}^0\odot\mathbf{W}^0 + s\{(\mathbf{W}\odot\mathbf{W}) - (\mathbf{W}^0\odot\mathbf{W}^0)\}}\left[(\mathbf{W}\odot\mathbf{W}) - (\mathbf{W}^0\odot\mathbf{W}^0)\right] ds\right\|_F \\
&\le \int_0^1 \exp\left(\left\|\mathbf{W}^0\odot\mathbf{W}^0 + s\{(\mathbf{W}\odot\mathbf{W}) - (\mathbf{W}^0\odot\mathbf{W}^0)\}\right\|_2\right) ds \\
&\qquad \times \|(\mathbf{W}\odot\mathbf{W}) - (\mathbf{W}^0\odot\mathbf{W}^0)\|_F \\
&\le e^{(\|\mathbf{W}^0\|_F + r)^2}(2\|\mathbf{W}^0\|_F + r)\|\mathbf{W} - \mathbf{W}^0\|_F,
\end{aligned}$$

we can get

$$\begin{aligned}|h(\mathbf{W}) - h(\mathbf{W}^0)| &= \left|\mathrm{tr}\Big(e^{\mathbf{W}\odot\mathbf{W}} - e^{\mathbf{W}^0\odot\mathbf{W}^0}\Big)\right| \\ &\le \sqrt{d}\|e^{\mathbf{W}\odot\mathbf{W}} - e^{\mathbf{W}^0\odot\mathbf{W}^0}\|_F \\ &\le \sqrt{d}\, e^{(\|\mathbf{W}^0\|_F+r)^2}(2\|\mathbf{W}^0\|_F + r)\|\mathbf{W} - \mathbf{W}^0\|_F.\end{aligned}$$

Likewise,

$$\begin{aligned}&\left\|2\lambda_2\mathbf{W}\odot(e^{\mathbf{W}\odot\mathbf{W}})^\top - 2\lambda_2^0\mathbf{W}^0\odot(e^{\mathbf{W}^0\odot\mathbf{W}^0})^\top\right\|_F \\ &\le 2|\lambda_2-\lambda_2^0|\|\mathbf{W}\odot(e^{\mathbf{W}\odot\mathbf{W}})^\top\|_F + 2\lambda_2^0\|(\mathbf{W}-\mathbf{W}^0)\odot(e^{\mathbf{W}\odot\mathbf{W}})^\top\|_F + 2\lambda_2^0\|\mathbf{W}^0\odot(e^{\mathbf{W}\odot\mathbf{W}} - e^{\mathbf{W}^0\odot\mathbf{W}^0})^\top\|_F \\ &\le 2|\lambda_2-\lambda_2^0|\|\mathbf{W}\|_F\|e^{\mathbf{W}\odot\mathbf{W}}\|_F + 2\lambda_2^0\|\mathbf{W}-\mathbf{W}^0\|_F\|e^{\mathbf{W}\odot\mathbf{W}}\|_F + 2\lambda_2^0\|\mathbf{W}^0\|_F\|e^{\mathbf{W}\odot\mathbf{W}} - e^{\mathbf{W}^0\odot\mathbf{W}^0}\|_F \\ &\le 2\sqrt{d}\,|\lambda_2-\lambda_2^0|(\|\mathbf{W}^0\|_F + r)e^{(\|\mathbf{W}^0\|_F+r)^2} + 2\sqrt{d}\,\lambda_2^0 e^{(\|\mathbf{W}^0\|_F+r)^2}\|\mathbf{W}-\mathbf{W}^0\|_F \\ &\qquad + 2\lambda_2^0\|\mathbf{W}^0\|_F e^{(\|\mathbf{W}^0\|_F+r)^2}(2\|\mathbf{W}^0\|_F + r)\|\mathbf{W}-\mathbf{W}^0\|_F.\end{aligned}$$

The residual-gradient perturbation obeys

$$\begin{aligned}&\left\|-d\frac{\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + d\frac{\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}\right\|_F \\ &\le d\left\|\frac{\mathbf{X}^\top\mathbf{X}(\mathbf{W}^0-\mathbf{W})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right\|_F + d\left\|\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\left(\frac{1}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{1}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}\right)\right\|_F \\ &\le \frac{2d\|\mathbf{X}^\top\mathbf{X}\|_2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}\|\mathbf{W}-\mathbf{W}^0\|_F + \frac{2d\|\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\|_F}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^4}\left|\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2\right| \\ &\le \left[\frac{2d\|\mathbf{X}^\top\mathbf{X}\|_2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2} + \frac{2d\|\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\|_F\left(2\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F\|\mathbf{X}\|_2 + r\|\mathbf{X}\|_2^2\right)}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^4}\right]\|\mathbf{W}-\mathbf{W}^0\|_F.\end{aligned}$$

The inverse-gradient perturbation obeys

$$\begin{aligned}\|(\mathbf{I}-\mathbf{W})^{-\top} - (\mathbf{I}-\mathbf{W}^0)^{-\top}\|_F &= \left\|(\mathbf{I}-\mathbf{W})^{-\top}(\mathbf{W}-\mathbf{W}^0)^\top(\mathbf{I}-\mathbf{W}^0)^{-\top}\right\|_F \\ &\le \|(\mathbf{I}-\mathbf{W})^{-1}\|_2\|(\mathbf{I}-\mathbf{W}^0)^{-1}\|_2\|\mathbf{W}-\mathbf{W}^0\|_F \\ &\le \frac{\|(\mathbf{I}-\mathbf{W}^0)^{-1}\|_2^2}{1-\|(\mathbf{I}-\mathbf{W}^0)^{-1}\|_2\|\mathbf{W}-\mathbf{W}^0\|_F}\|\mathbf{W}-\mathbf{W}^0\|_F.\end{aligned}$$

These estimates show not only continuity but locally uniform control of the complete GOLEM stationarity field, with the acyclicity term treated at the same analytic order as the residual likelihood and the log-Jacobian term. Now consider the active equations

$$0 = -d\frac{\big[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\big]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + \big[(\mathbf{I}-\mathbf{W})^{-\top}\big]_{ij} + 2\lambda_2[\mathbf{W}]_{ij}\big[e^{\mathbf{W}\odot\mathbf{W}}\big]_{ji} + \lambda_1\,\mathrm{sgn}([\mathbf{W}^0]_{ij}), \quad (i,j)\in\mathcal{A},$$

together with the active-stratum equations $[\mathbf{W}]_{ij} = 0$ for $(i,j)\notin\mathcal{A}$. The active system is real analytic in the active entries of $\mathbf{W}$ and in (λ_1,λ_2) on the domain where $\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 > 0$ and $\det(\mathbf{I}-\mathbf{W})\neq 0$. At $(\mathbf{W}^0,\lambda_1^0,\lambda_2^0)$, its Jacobian with respect to the active variables $[\mathbf{W}]_{kl}$, $(k,l)\in\mathcal{A}$, is precisely

$$\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \log|\det(\mathbf{I}-\mathbf{W})| + \lambda_2^0 h(\mathbf{W})\right\}\Big|_{\mathbf{W}=\mathbf{W}^0}\right]_{(i,j),(k,l)\in\mathcal{A}},$$

which is nonsingular by assumption. Therefore the real-analytic implicit function theorem gives real-analytic functions $(\lambda_1,\lambda_2)\mapsto[\mathbf{W}(\lambda_1,\lambda_2)]_{ij}$, $(i,j)\in\mathcal{A}$, near $(\lambda_1^0,\lambda_2^0)$, satisfying $[\mathbf{W}(\lambda_1^0,\lambda_2^0)]_{ij} = [\mathbf{W}^0]_{ij}$, $(i,j)\in\mathcal{A}$, and the active stationarity equations. Define $[\mathbf{W}(\lambda_1,\lambda_2)]_{ij} = 0$ for $(i,j)\notin\mathcal{A}$. This gives a real-analytic matrix-valued map before the final shrinking of the parameter neighborhood.

We next formally prove that the analytic active-stratum solution is a full nonsmooth stationary point, not merely a solution of the restricted equations. Since $(\lambda_1^0,\lambda_2^0)$ is interior to $(0,\lambda_{\max})\times(0,\infty)$,

since $\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2>0$ and $\det(\mathbf{I}-\mathbf{W}^0)\neq 0$, and since the inactive inequalities are strict, choose $\varepsilon>0$ so that $\mathbb{B}_\varepsilon((\lambda_1^0,\lambda_2^0))\subset(0,\lambda_{\max})\times(0,\infty)$, we can obtain

$$\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2>\frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2,\qquad |\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))|>\frac{1}{2}|\det(\mathbf{I}-\mathbf{W}^0)|,$$

$\big|[\mathbf{W}(\lambda_1,\lambda_2)]_{ij}-[\mathbf{W}^0]_{ij}\big|<\frac{1}{2}|[\mathbf{W}^0]_{ij}|$ for $(i,j)\in\mathcal{A}$, and, for every $(i,j)\notin\mathcal{A}$,

$$\begin{aligned}
&\left|\left[-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2}+\left[(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{ij}+2\lambda_2[\mathbf{W}(\lambda_1,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}\right]\right.\\
&\qquad\left.-\left[-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right]\right|\\
&<\frac{1}{5}\left[\lambda_1^0-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\right],
\end{aligned}$$

and

$$|\lambda_1-\lambda_1^0|<\frac{1}{5}\left[\lambda_1^0-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\right].$$

Then, for $(i,j)\in\mathcal{A}$, $\operatorname{sgn}([\mathbf{W}(\lambda_1,\lambda_2)]_{ij})=\operatorname{sgn}([\mathbf{W}^0]_{ij})$, and the active equations give

$$\begin{aligned}
0=&-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2}+\left[(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{ij}\\
&+2\lambda_2[\mathbf{W}(\lambda_1,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}+\lambda_1\operatorname{sgn}([\mathbf{W}(\lambda_1,\lambda_2)]_{ij}).
\end{aligned}$$

For $(i,j)\notin\mathcal{A}$, the strict inequality is preserved by the chain

$$\begin{aligned}
&\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2}+\left[(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{ij}+2\lambda_2[\mathbf{W}(\lambda_1,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}\right|\\
&\le\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\\
&\quad+\frac{1}{5}\left[\lambda_1^0-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\right]\\
&=\lambda_1^0-\frac{4}{5}\left[\lambda_1^0-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\right]\\
&<\lambda_1^0-\frac{1}{5}\left[\lambda_1^0-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}^0)\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}^0\|_F^2}+\left[(\mathbf{I}-\mathbf{W}^0)^{-\top}\right]_{ij}+2\lambda_2^0[\mathbf{W}^0]_{ij}\left[e^{\mathbf{W}^0\odot\mathbf{W}^0}\right]_{ji}\right|\right]\\
&\le\lambda_1.
\end{aligned}$$

Thus

$$\mathbf{0}\in\nabla_{\mathbf{W}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))|+\lambda_2 h(\mathbf{W}(\lambda_1,\lambda_2))\right\}+\lambda_1\partial\|\mathbf{W}(\lambda_1,\lambda_2)\|_1,$$

so $\mathbf{W}(\lambda_1,\lambda_2)$ is a stationary point of (3) throughout $\mathbb{B}_\varepsilon((\lambda_1^0,\lambda_2^0))$. Next, we derive the PDE system for the two-parameter solution surface. Differentiating the active stationarity equations with respect

to λ_1 yields, for every $(i,j) \in \mathcal{A}$,

$$
\begin{aligned}
0 = \sum_{(k,l)\in\mathcal{A}} \Bigg[& d \frac{\left[\mathbf{X}^\top \mathbf{X}\right]_{ik} \mathbf{1}_{\{j=l\}}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2} - 2d \frac{\left[\mathbf{X}^\top (\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{ij} \left[\mathbf{X}^\top (\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{kl}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^4} \\
& + 2\lambda_2 \mathbf{1}_{\{i=k\}} \mathbf{1}_{\{j=l\}} \left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji} + \left[(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{il} \left[(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{kj} \\
& + 4\lambda_2 [\mathbf{W}(\lambda_1,\lambda_2)]_{ij} [\mathbf{W}(\lambda_1,\lambda_2)]_{kl} \int_0^1 \left[e^{(1-s)(\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2))}\right]_{jk} \\
& \qquad \times \left[e^{s(\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2))}\right]_{li} ds \Bigg] \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{kl}}{\partial \lambda_1} + \operatorname{sgn}([\mathbf{W}^0]_{ij}).
\end{aligned} \tag{27}
$$

Differentiating with respect to λ_2 yields, again for every $(i,j) \in \mathcal{A}$,

$$
\begin{aligned}
0 = \sum_{(k,l)\in\mathcal{A}} \Bigg[& d \frac{\left[\mathbf{X}^\top \mathbf{X}\right]_{ik} \mathbf{1}_{\{j=l\}}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2} - 2d \frac{\left[\mathbf{X}^\top (\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{ij} \left[\mathbf{X}^\top (\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2))\right]_{kl}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^4} \\
& + 2\lambda_2 \mathbf{1}_{\{i=k\}} \mathbf{1}_{\{j=l\}} \left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji} + \left[(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{il} \left[(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2))^{-\top}\right]_{kj} \\
& + 4\lambda_2 [\mathbf{W}(\lambda_1,\lambda_2)]_{ij} [\mathbf{W}(\lambda_1,\lambda_2)]_{kl} \int_0^1 \left[e^{(1-s)(\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2))}\right]_{jk} \\
& \qquad \times \left[e^{s(\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2))}\right]_{li} ds \Bigg] \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{kl}}{\partial \lambda_2} \\
& + 2[\mathbf{W}(\lambda_1,\lambda_2)]_{ij} \left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}.
\end{aligned} \tag{28}
$$

For inactive entries, $\frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{ij}}{\partial \lambda_1} = \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{ij}}{\partial \lambda_2} = 0$ for $(i,j) \notin \mathcal{A}$. This is the GOLEM-specific first-order PDE system for the local stationary surface. Its coefficient matrix is exactly the active restriction of the full residual–log-determinant–matrix-exponential curvature, and its λ_2-forcing term is the acyclicity gradient $2[\mathbf{W}(\lambda_1,\lambda_2)]_{ij} \left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}$. The PDE systems (27), (28) also characterize two-sided directional behavior. For arbitrary $a, b \in \mathbb{R}$,

$$
\begin{aligned}
& \lim_{t\downarrow 0} \frac{[\mathbf{W}(\lambda_1 + ta, \lambda_2 + tb)]_{ij} - [\mathbf{W}(\lambda_1,\lambda_2)]_{ij}}{t} \\
= & \lim_{t\downarrow 0} \frac{[\mathbf{W}(\lambda_1,\lambda_2)]_{ij} - [\mathbf{W}(\lambda_1 - ta, \lambda_2 - tb)]_{ij}}{t} \\
= & \, a \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{ij}}{\partial \lambda_1} + b \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{ij}}{\partial \lambda_2},
\end{aligned}
$$

and for active coordinates the common value is determined by

$$
\begin{aligned}
0 = \sum_{(k,l)\in\mathcal{A}} & \frac{\partial^2}{\partial [\mathbf{W}]_{ij} \partial [\mathbf{W}]_{kl}} \left\{ \frac{d}{2} \log \|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2 - \log|\det(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2))| + \lambda_2 h(\mathbf{W}(\lambda_1,\lambda_2)) \right\} \\
& \times \left[a \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{kl}}{\partial \lambda_1} + b \frac{\partial [\mathbf{W}(\lambda_1,\lambda_2)]_{kl}}{\partial \lambda_2} \right] + a \, \operatorname{sgn}([\mathbf{W}^0]_{ij}) + 2b \, [\mathbf{W}(\lambda_1,\lambda_2)]_{ij} \left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{ji}.
\end{aligned}
$$

For inactive coordinates this common directional derivative is zero. Hence the solution surface has matching left and right directional derivatives in every parameter direction inside the transition-free neighborhood. Restricting the PDE to coordinate lines gives the exact coordinate-wise ODEs. Along $\lambda_1 \mapsto \mathbf{W}(\lambda_1, \lambda_2^0)$, we have

$$
\begin{aligned}
0 = \sum_{(k,l)\in\mathcal{A}} & \frac{\partial^2}{\partial [\mathbf{W}]_{ij} \partial [\mathbf{W}]_{kl}} \left\{ \frac{d}{2} \log \|\mathbf{X} - \mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0)\|_F^2 - \log|\det(\mathbf{I} - \mathbf{W}(\lambda_1,\lambda_2^0))| + \lambda_2^0 h(\mathbf{W}(\lambda_1,\lambda_2^0)) \right\} \frac{d[\mathbf{W}(\lambda_1,\lambda_2^0)]_{kl}}{d\lambda_1} \\
& + \operatorname{sgn}([\mathbf{W}^0]_{ij}), \qquad (i,j) \in \mathcal{A},
\end{aligned}
$$

and $\frac{d[\mathbf{W}(\lambda_1,\lambda_2^0)]_{ij}}{d\lambda_1}=0$ for $(i,j)\notin\mathcal{A}$. Along $\lambda_2\mapsto\mathbf{W}(\lambda_1^0,\lambda_2)$,

$$0=\sum_{(k,l)\in\mathcal{A}}\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1^0,\lambda_2))|+\lambda_2 h(\mathbf{W}(\lambda_1^0,\lambda_2))\right\}\frac{d[\mathbf{W}(\lambda_1^0,\lambda_2)]_{kl}}{d\lambda_2}$$
$$+2[\mathbf{W}(\lambda_1^0,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1^0,\lambda_2)\odot\mathbf{W}(\lambda_1^0,\lambda_2)}\right]_{ji},\qquad(i,j)\in\mathcal{A},$$

and $\frac{d[\mathbf{W}(\lambda_1^0,\lambda_2)]_{ij}}{d\lambda_2}=0$ for $(i,j)\notin\mathcal{A}$. Thus the λ_1-ODE is forced by the fixed ℓ_1-sign vector, while the λ_2-ODE is forced by the exact gradient of h, and both are preconditioned by the same GOLEM curvature containing the residual likelihood term, the log-determinant term, and the noncommutative exponential-integral curvature of h. The coupling to the original objective is exact. Along the λ_1-path,

$$\frac{d}{d\lambda_1}\left[\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2^0))|+\lambda_1\|\mathbf{W}(\lambda_1,\lambda_2^0)\|_1+\lambda_2^0 h(\mathbf{W}(\lambda_1,\lambda_2^0))\right]$$
$$=\sum_{(i,j)\in\mathcal{A}}\left[-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0))\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0)\|_F^2}+\left[(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2^0))^{-\top}\right]_{ij}\right.$$
$$\left.+2\lambda_2^0[\mathbf{W}(\lambda_1,\lambda_2^0)]_{ij}\left[e^{\mathbf{W}(\lambda_1,\lambda_2^0)\odot\mathbf{W}(\lambda_1,\lambda_2^0)}\right]_{ji}+\lambda_1\operatorname{sgn}([\mathbf{W}^0]_{ij})\right]\frac{d[\mathbf{W}(\lambda_1,\lambda_2^0)]_{ij}}{d\lambda_1}+\|\mathbf{W}(\lambda_1,\lambda_2^0)\|_1$$
$$=\|\mathbf{W}(\lambda_1,\lambda_2^0)\|_1.$$

Along the λ_2-path,

$$\frac{d}{d\lambda_2}\left[\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1^0,\lambda_2))|+\lambda_1^0\|\mathbf{W}(\lambda_1^0,\lambda_2)\|_1+\lambda_2 h(\mathbf{W}(\lambda_1^0,\lambda_2))\right]$$
$$=\sum_{(i,j)\in\mathcal{A}}\left[-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2))\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2)\|_F^2}+\left[(\mathbf{I}-\mathbf{W}(\lambda_1^0,\lambda_2))^{-\top}\right]_{ij}\right.$$
$$\left.+2\lambda_2[\mathbf{W}(\lambda_1^0,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1^0,\lambda_2)\odot\mathbf{W}(\lambda_1^0,\lambda_2)}\right]_{ji}+\lambda_1^0\operatorname{sgn}([\mathbf{W}^0]_{ij})\right]\frac{d[\mathbf{W}(\lambda_1^0,\lambda_2)]_{ij}}{d\lambda_2}+h(\mathbf{W}(\lambda_1^0,\lambda_2))$$
$$=h(\mathbf{W}(\lambda_1^0,\lambda_2)).$$

The second-order envelope identities are correspondingly

$$\frac{d^2}{d\lambda_1^2}\left[\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2^0))|+\lambda_1\|\mathbf{W}(\lambda_1,\lambda_2^0)\|_1+\lambda_2^0 h(\mathbf{W}(\lambda_1,\lambda_2^0))\right]$$
$$=-\sum_{(i,j)\in\mathcal{A}}\sum_{(k,l)\in\mathcal{A}}\operatorname{sgn}([\mathbf{W}^0]_{ij})\times$$
$$\left(\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2^0)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2^0))|+\lambda_2^0 h(\mathbf{W}(\lambda_1,\lambda_2^0))\right\}\right]^{-1}_{(i,j),(k,l)\in\mathcal{A}}\right)\operatorname{sgn}([\mathbf{W}^0]_{kl}),\tag{29}$$

along with

$$\frac{d^2}{d\lambda_2^2}\left[\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1^0,\lambda_2))|+\lambda_1^0\|\mathbf{W}(\lambda_1^0,\lambda_2)\|_1+\lambda_2 h(\mathbf{W}(\lambda_1^0,\lambda_2))\right]$$
$$=-\sum_{(i,j)\in\mathcal{A}}\sum_{(k,l)\in\mathcal{A}}2[\mathbf{W}(\lambda_1^0,\lambda_2)]_{ij}\left[e^{\mathbf{W}(\lambda_1^0,\lambda_2)\odot\mathbf{W}(\lambda_1^0,\lambda_2)}\right]_{ji}$$
$$\times\left(\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1^0,\lambda_2))|+\lambda_2 h(\mathbf{W}(\lambda_1^0,\lambda_2))\right\}\right]^{-1}_{(i,j),(k,l)\in\mathcal{A}}\right)$$
$$\times 2[\mathbf{W}(\lambda_1^0,\lambda_2)]_{kl}\left[e^{\mathbf{W}(\lambda_1^0,\lambda_2)\odot\mathbf{W}(\lambda_1^0,\lambda_2)}\right]_{lk}.\tag{30}$$

The mixed envelope curvature is

$$\begin{aligned}
&\frac{\partial^2}{\partial\lambda_1\partial\lambda_2}\left[\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))|+\lambda_1\|\mathbf{W}(\lambda_1,\lambda_2)\|_1+\lambda_2 h(\mathbf{W}(\lambda_1,\lambda_2))\right]\\
&=-\sum_{(i,j)\in\mathcal{A}}\sum_{(k,l)\in\mathcal{A}}\operatorname{sgn}([\mathbf{W}^0]_{ij})\times\\
&\quad\left(\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1,\lambda_2)\|_F^2-\log|\det(\mathbf{I}-\mathbf{W}(\lambda_1,\lambda_2))|+\lambda_2 h(\mathbf{W}(\lambda_1,\lambda_2))\right\}\right]^{-1}_{(i,j),(k,l)\in\mathcal{A}}\right)\\
&\qquad\times 2[\mathbf{W}(\lambda_1,\lambda_2)]_{kl}\left[e^{\mathbf{W}(\lambda_1,\lambda_2)\odot\mathbf{W}(\lambda_1,\lambda_2)}\right]_{lk}.
\end{aligned}\tag{31}$$

These identities show that h enters simultaneously as the explicit λ_2-sensitivity, as the forcing field for the λ_2-flow, and as a noncommutative exponential-integral contribution to the curvature through which both λ_1- and λ_2-directions are propagated. Since $(\lambda_1,\lambda_2)\mapsto\mathbf{W}(\lambda_1,\lambda_2)$ is real analytic, the coordinate restrictions $\lambda_1\mapsto\mathbf{W}(\lambda_1,\lambda_2^0)$ and $\lambda_2\mapsto\mathbf{W}(\lambda_1^0,\lambda_2)$ are real analytic, hence C^∞, in neighborhoods of λ_1^0 and λ_2^0, respectively.

It remains to prove analytic continuation. Let a connected transition-free subset of $(0,\lambda_{\max})\times(0,\infty)$ be fixed, and let a continuous piecewise C^1 path inside it begin at $(\lambda_1^0,\lambda_2^0)$. At every lifted point reached before transition, the same conditions hold: $[\mathbf{W}]_{ij}\neq 0$ for $(i,j)\in\mathcal{A}$, and

$$\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}+2\lambda_2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}\right|<\lambda_1,\qquad(i,j)\notin\mathcal{A},$$

$$\det\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2-\log|\det(\mathbf{I}-\mathbf{W})|+\lambda_2 h(\mathbf{W})\right\}\right]_{(i,j),(k,l)\in\mathcal{A}}\neq 0,$$

with $\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2>0$ and $\det(\mathbf{I}-\mathbf{W})\neq 0$. Therefore the same real-analytic implicit-function construction applies at each lifted point. On compact subpaths, finitely many such analytic charts cover the branch. On overlaps, uniqueness in the analytic implicit-function theorem forces the charts to agree (27)-(28), since they solve the same active stationarity equations, impose the same inactive zero pattern, and retain the same signs. Hence the local analytic surface can be continued by overlapping analytic charts along the path. The continuation of (29)-(31) can fail only at the boundary of the transition-free regime, namely only if one of the following quantities degenerates:
$\min_{(i,j)\in\mathcal{A}}|[\mathbf{W}]_{ij}|\downarrow 0$,

$$\min_{(i,j)\notin\mathcal{A}}\left\{\lambda_1-\left|-d\frac{\left[\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{W})\right]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+\left[(\mathbf{I}-\mathbf{W})^{-\top}\right]_{ij}+2\lambda_2[\mathbf{W}]_{ij}\left[e^{\mathbf{W}\odot\mathbf{W}}\right]_{ji}\right|\right\}\downarrow 0,$$

$$\det\left[\frac{\partial^2}{\partial[\mathbf{W}]_{ij}\partial[\mathbf{W}]_{kl}}\left\{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2-\log|\det(\mathbf{I}-\mathbf{W})|+\lambda_2 h(\mathbf{W})\right\}\right]_{(i,j),(k,l)\in\mathcal{A}}\to 0,$$

or

$$\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2\downarrow 0,\qquad\det(\mathbf{I}-\mathbf{W})\to 0,\qquad\lambda_1\downarrow 0,\qquad\lambda_1\uparrow\lambda_{\max},\qquad\lambda_2\downarrow 0,$$

or $\mathbf{W}$ escapes every bounded subset of $\mathbb{R}^{d\times d}$. These are precisely the support-loss, inactive-entry, curvature-degeneracy, smooth-domain, admissible-boundary, and escape transition mechanisms. Therefore the local surface admits analytic continuation along any connected transition-free subset of the admissible parameter region, up to the nearest transition locus. □

I Proof of Theorem 4.3

Proof. Fix a transition-free interval. Then $\mathbf{W}_{\mathcal{A}^c}\equiv\mathbf{0}$, the active signs $\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j})$ are constant for every j, $\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2>0$, and $\det(\mathbf{I}-\mathbf{W})\neq 0$. For use inside this proof only, define the active stationary residual

$$\mathcal{F}(\mathbf{W}_{\mathcal{A}},\lambda_1,\lambda_2)=\left(\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-\top}]_{\mathcal{A}_j,j}+2\lambda_2\mathbf{W}_{\mathcal{A}_j,j}\odot[(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}+\lambda_1\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j})\right)_{j=1}^{d}.$$

Here $\mathcal{F}$ is viewed as a map on the finite-dimensional active coordinate space $\{\mathbf{U} \in \mathbb{R}^{d\times d} : \mathbf{U}_{\mathcal{A}^c} = \mathbf{0}\}$, with the convention that the j-th block is void when $\mathcal{A}_j = \varnothing$. We first verify the active stationarity condition of (3) through difference quotients. For every active perturbation $\mathbf{U}_{\mathcal{A}^c} = \mathbf{0}$, the profiled residual term satisfies

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{\frac{d}{2}\log\|\mathbf{X}-\mathbf{X}(\mathbf{W}+\epsilon\mathbf{U})\|_F^2 - \frac{d}{2}\log\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{d}{2} \frac{\log\left(1+\frac{-2\epsilon\langle \mathbf{X}-\mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F + \epsilon^2\|\mathbf{X}\mathbf{U}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right)}{\epsilon} \\
&= \frac{d}{2}\lim_{\epsilon\to 0}\left[\frac{-2\langle \mathbf{X}-\mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F + \epsilon\|\mathbf{X}\mathbf{U}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right]\left[\frac{\log\left(1+\epsilon\frac{-2\langle \mathbf{X}-\mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F + \epsilon\|\mathbf{X}\mathbf{U}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right)}{\epsilon\frac{-2\langle \mathbf{X}-\mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F + \epsilon\|\mathbf{X}\mathbf{U}\|_F^2}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}}\right] \\
&= -\frac{d\langle \mathbf{X}-\mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} = \left\langle \frac{d\,\mathbf{X}^\top\mathbf{X}(\mathbf{W}-\mathbf{I})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}, \mathbf{U}\right\rangle_F.
\end{aligned}$$

For the log-determinant term, for sufficiently small $|\epsilon|$, we have

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{-\log|\det(\mathbf{I}-\mathbf{W}-\epsilon\mathbf{U})| + \log|\det(\mathbf{I}-\mathbf{W})|}{\epsilon} \\
&= -\lim_{\epsilon\to 0} \frac{\log\left|\det\left(\mathbf{I}-\epsilon(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}\right)\right|}{\epsilon} \\
&= -\lim_{\epsilon\to 0} \frac{\operatorname{tr}\log\left(\mathbf{I}-\epsilon(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}\right)}{\epsilon} \\
&= -\lim_{\epsilon\to 0} \frac{\operatorname{tr}\left[-\epsilon(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U} - \frac{\epsilon^2}{2}\{(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}\}^2 + \epsilon^2\left\{\frac{\log(\mathbf{I}-\epsilon(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}) + \epsilon(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U} + \frac{\epsilon^2}{2}\{(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}\}^2}{\epsilon^2}\right\}\right]}{\epsilon} \\
&= \operatorname{tr}\{(\mathbf{I}-\mathbf{W})^{-1}\mathbf{U}\} = \left\langle (\mathbf{I}-\mathbf{W})^{-\top}, \mathbf{U}\right\rangle_F.
\end{aligned}$$

For the analytic penalty $h(\mathbf{W})$, use the finite exponential increment identity $e^{\mathbf{B}+\mathbf{C}} - e^{\mathbf{B}} = \int_0^1 e^{(1-s)(\mathbf{B}+\mathbf{C})}\mathbf{C}e^{s\mathbf{B}}\,ds$, which follows from

$$\begin{aligned}
e^{\mathbf{B}+\mathbf{C}} - e^{\mathbf{B}} &= -\int_0^1 \frac{d}{ds}\left[e^{(1-s)(\mathbf{B}+\mathbf{C})}e^{s\mathbf{B}}\right] ds \\
&= -\int_0^1 \left[-e^{(1-s)(\mathbf{B}+\mathbf{C})}(\mathbf{B}+\mathbf{C})e^{s\mathbf{B}} + e^{(1-s)(\mathbf{B}+\mathbf{C})}\mathbf{B}e^{s\mathbf{B}}\right] ds = \int_0^1 e^{(1-s)(\mathbf{B}+\mathbf{C})}\mathbf{C}e^{s\mathbf{B}}\,ds.
\end{aligned}$$

Substituting $\mathbf{B} = \mathbf{W}\odot\mathbf{W}$ and $\mathbf{C} = 2\epsilon\mathbf{W}\odot\mathbf{U} + \epsilon^2\mathbf{U}\odot\mathbf{U}$ gives

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{h(\mathbf{W}+\epsilon\mathbf{U})-h(\mathbf{W})}{\epsilon} &= \lim_{\epsilon\to 0}\operatorname{tr}\left[\int_0^1 e^{(1-s)\{\mathbf{W}\odot\mathbf{W}+2\epsilon\mathbf{W}\odot\mathbf{U}+\epsilon^2\mathbf{U}\odot\mathbf{U}\}}(2\mathbf{W}\odot\mathbf{U}+\epsilon\mathbf{U}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds\right] \\
&= \operatorname{tr}\left[\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds\right] \\
&= \int_0^1 \operatorname{tr}\left[e^{s(\mathbf{W}\odot\mathbf{W})}e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})\right] ds \\
&= \operatorname{tr}\left[e^{\mathbf{W}\odot\mathbf{W}}(2\mathbf{W}\odot\mathbf{U})\right] = \left\langle 2\mathbf{W}\odot(e^{\mathbf{W}\odot\mathbf{W}})^\top, \mathbf{U}\right\rangle_F.
\end{aligned}$$

Finally, on the fixed active orthant, $\|\mathbf{W}+\epsilon\mathbf{U}\|_1 - \|\mathbf{W}\|_1 = \epsilon\sum_{j=1}^d \operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j})^\top\mathbf{U}_{\mathcal{A}_j,j}$ for all sufficiently small $|\epsilon|$. Therefore the first-order stationarity of (3) on the active coordinates is precisely

$$\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-\top}]_{\mathcal{A}_j,j} + 2\lambda_2\mathbf{W}_{\mathcal{A}_j,j}\odot[(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} + \lambda_1\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j}) = \mathbf{0}, \quad j = 1,\dots,d.$$

We next verify the regularity needed for the implicit-function step. Fix any point on the transition-free branch. Choose $\rho \in (0,1]$ so that, whenever $\mathbf{U}_{\mathcal{A}^c} = \mathbf{0}$ and $\|\mathbf{U}\|_F \le \rho$, $\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2 > 0$ and $\det(\mathbf{I} - \mathbf{W} - \mathbf{U}) \neq 0$. Define the finite local bound

$$M := 1 + \|\mathbf{X}\|_{\mathrm{op}} + \|\mathbf{X}\|_F + \sup_{\substack{\mathbf{U}_{\mathcal{A}^c}=\mathbf{0} \\ \|\mathbf{U}\|_F \le \rho}} \|\mathbf{W}+\mathbf{U}\|_F + \sup_{\substack{\mathbf{U}_{\mathcal{A}^c}=\mathbf{0} \\ \|\mathbf{U}\|_F \le \rho}} \|(\mathbf{I}-\mathbf{W}-\mathbf{U})^{-\top}\|_{\mathrm{op}} + \sup_{\substack{\mathbf{U}_{\mathcal{A}^c}=\mathbf{0} \\ \|\mathbf{U}\|_F \le \rho}} \|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^{-2}.$$

Then $M < \infty$, $M \ge 1$, and all subsequent bounds are expressed explicitly in terms of this M. The residual scale has the exact form

$$\begin{aligned}\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 &= \mathrm{tr}\left[\{\mathbf{X} - \mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{U}\}^\top \{\mathbf{X} - \mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{U}\} - (\mathbf{X} - \mathbf{X}\mathbf{W})^\top (\mathbf{X} - \mathbf{X}\mathbf{W})\right] \\ &= -2\langle \mathbf{X} - \mathbf{X}\mathbf{W}, \mathbf{X}\mathbf{U}\rangle_F + \|\mathbf{X}\mathbf{U}\|_F^2 \\ &= 2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F + \|\mathbf{X}\mathbf{U}\|_F^2.\end{aligned}$$

Since $\|\mathbf{X}\mathbf{W} - \mathbf{X}\|_F \le \|\mathbf{X}\|_{\mathrm{op}}\|\mathbf{W}\|_F + \|\mathbf{X}\|_F \le M^2 + M \le 2M^2$ and $\|\mathbf{X}\mathbf{U}\|_F \le M\|\mathbf{U}\|_F$, we have

$$\begin{aligned}\left|2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F + \|\mathbf{X}\mathbf{U}\|_F^2\right| &\le 2\|\mathbf{X}\mathbf{W} - \mathbf{X}\|_F\|\mathbf{X}\mathbf{U}\|_F + \|\mathbf{X}\mathbf{U}\|_F^2 \\ &\le 4M^3\|\mathbf{U}\|_F + M^2\|\mathbf{U}\|_F^2 \\ &\le 5M^3\|\mathbf{U}\|_F.\end{aligned}$$

The reciprocal residual map therefore satisfies the second-order estimate

$$\begin{aligned}&\left|\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^{-2} - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^{-2} + \frac{2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\right| \\ &= \left|-\frac{2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F + \|\mathbf{X}\mathbf{U}\|_F^2}{\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + \frac{2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\right| \\ &= \left|\frac{2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F \left[\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2\right] - \|\mathbf{X}\mathbf{U}\|_F^2\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}{\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\right| \\ &\le 2M^3\left|\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F\right|\left|\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2\right| + M^2\|\mathbf{X}\mathbf{U}\|_F^2 \\ &\le 2M^3(2M^3\|\mathbf{U}\|_F)(5M^3\|\mathbf{U}\|_F) + M^2(M^2\|\mathbf{U}\|_F^2) \\ &\le (20M^9 + M^4)\|\mathbf{U}\|_F^2 \\ &\le 21M^9\|\mathbf{U}\|_F^2.\end{aligned}$$

This bound is used to control the coupled quotient in the profiled likelihood. Now for every j,

$$\begin{aligned}&\Bigg\| d\,\frac{\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}(\mathbf{W}+\mathbf{U})_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} \\ &\quad - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top\mathbf{X}_{:,\mathcal{A}_j}\mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\sum_{r=1}^d \left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})\right]^\top \mathbf{U}_{\mathcal{A}_r,r}\Bigg\|_2 \\ &\le d\,\|\mathbf{X}_{:,\mathcal{A}_j}^\top\mathbf{X}_{:,\mathcal{A}_j}\mathbf{U}_{\mathcal{A}_j,j}\|_2\left|\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^{-2} - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^{-2}\right| \\ &\quad + d\,\|\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})\|_2\left|\|\mathbf{X} - \mathbf{X}(\mathbf{W}+\mathbf{U})\|_F^{-2} - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^{-2} + \frac{2\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\right| \\ &\quad + \frac{2d\,\|\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})\|_2}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4}\left|\langle \mathbf{X}\mathbf{W} - \mathbf{X}, \mathbf{X}\mathbf{U}\rangle_F - \sum_{r=1}^d \left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})\right]^\top \mathbf{U}_{\mathcal{A}_r,r}\right| \\ &\le d(M^2\|\mathbf{U}\|_F)(5M^5\|\mathbf{U}\|_F) + d(M\cdot 2M^2)(21M^9\|\mathbf{U}\|_F^2) + 0 \\ &\le (5dM^7 + 42dM^{12})\|\mathbf{U}\|_F^2 \\ &\le 47dM^{12}\|\mathbf{U}\|_F^2.\end{aligned}$$

The zero in the third term follows from $\mathbf{U}_{\mathcal{A}^c} = \mathbf{0}$, because

$$\langle \mathbf{XW} - \mathbf{X}, \mathbf{XU} \rangle_F = \sum_{r=1}^{d} (\mathbf{XW}_{:,r} - \mathbf{X}_{:,r})^\top \mathbf{XU}_{:,r} = \sum_{r=1}^{d} \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{XW}_{:,r} - \mathbf{X}_{:,r}) \right]^\top \mathbf{U}_{\mathcal{A}_r,r}.$$

The inverse-transpose term has the exact resolvent remainder

$$\begin{aligned}
&(\mathbf{I} - \mathbf{W} - \mathbf{U})^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top} \mathbf{U}^\top (\mathbf{I} - \mathbf{W})^{-\top} \\
&= \left[(\mathbf{I} - \mathbf{W} - \mathbf{U})^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top} \right] \mathbf{U}^\top (\mathbf{I} - \mathbf{W})^{-\top} \\
&= (\mathbf{I} - \mathbf{W} - \mathbf{U})^{-\top} \mathbf{U}^\top (\mathbf{I} - \mathbf{W})^{-\top} \mathbf{U}^\top (\mathbf{I} - \mathbf{W})^{-\top},
\end{aligned}$$

and hence

$$\begin{aligned}
&\left\| (\mathbf{I} - \mathbf{W} - \mathbf{U})^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top} \mathbf{U}^\top (\mathbf{I} - \mathbf{W})^{-\top} \right\|_F \\
&\le \|(\mathbf{I} - \mathbf{W} - \mathbf{U})^{-\top}\|_{\mathrm{op}} \|\mathbf{U}\|_F \|(\mathbf{I} - \mathbf{W})^{-\top}\|_{\mathrm{op}} \|\mathbf{U}\|_F \|(\mathbf{I} - \mathbf{W})^{-\top}\|_{\mathrm{op}} \\
&\le M^3 \|\mathbf{U}\|_F^2.
\end{aligned}$$

For the exponential term, first note that $\|(\mathbf{W} + \mathbf{U}) \odot (\mathbf{W} + \mathbf{U})\|_{\mathrm{op}} \le \|(\mathbf{W} + \mathbf{U}) \odot (\mathbf{W} + \mathbf{U})\|_F \le \|\mathbf{W} + \mathbf{U}\|_F^2 \le M^2$, and therefore $\|e^{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})}\|_{\mathrm{op}} \le e^{M^2}$ and $\|e^{\mathbf{W} \odot \mathbf{W}}\|_{\mathrm{op}} \le e^{M^2}$. Moreover,

$$\begin{aligned}
&\|(\mathbf{W} + \mathbf{U}) \odot (\mathbf{W} + \mathbf{U}) - \mathbf{W} \odot \mathbf{W}\|_F \\
&= \|2\mathbf{W} \odot \mathbf{U} + \mathbf{U} \odot \mathbf{U}\|_F \\
&\le 2\|\mathbf{W}\|_F \|\mathbf{U}\|_F + \|\mathbf{U}\|_F^2 \\
&\le (2M + 1)\|\mathbf{U}\|_F \\
&\le 3M \|\mathbf{U}\|_F.
\end{aligned}$$

The second-order exponential remainder obeys

$$\begin{aligned}
&e^{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})} - e^{\mathbf{W} \odot \mathbf{W}} - \int_0^1 e^{(1-s)(\mathbf{W} \odot \mathbf{W})} (2\mathbf{W} \odot \mathbf{U}) e^{s(\mathbf{W} \odot \mathbf{W})} \, ds \\
&= \int_0^1 e^{(1-s)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} (\mathbf{U} \odot \mathbf{U}) e^{s(\mathbf{W} \odot \mathbf{W})} \, ds \\
&\quad + \int_0^1 \left[e^{(1-s)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} - e^{(1-s)(\mathbf{W} \odot \mathbf{W})} \right] (2\mathbf{W} \odot \mathbf{U}) e^{s(\mathbf{W} \odot \mathbf{W})} \, ds.
\end{aligned}$$

The first integral is bounded by

$$\begin{aligned}
\left\| \int_0^1 e^{(1-s)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} (\mathbf{U} \odot \mathbf{U}) e^{s(\mathbf{W} \odot \mathbf{W})} \, ds \right\|_F &\le \int_0^1 e^{(1-s)M^2} \|\mathbf{U} \odot \mathbf{U}\|_F e^{sM^2} \, ds \\
&\le e^{M^2} \|\mathbf{U}\|_F^2.
\end{aligned}$$

For the second integral,

$$\begin{aligned}
&\left\| e^{(1-s)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} - e^{(1-s)(\mathbf{W} \odot \mathbf{W})} \right\|_{\mathrm{op}} \\
&\le \int_0^{1-s} \left\| e^{(1-s-r)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} \left[2\mathbf{W} \odot \mathbf{U} + \mathbf{U} \odot \mathbf{U} \right] e^{r(\mathbf{W} \odot \mathbf{W})} \right\|_{\mathrm{op}} dr \\
&\le \int_0^{1-s} e^{(1-s-r)M^2} (3M \|\mathbf{U}\|_F) e^{rM^2} \, dr \\
&\le 3M e^{M^2} \|\mathbf{U}\|_F.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\left\| \int_0^1 \left[e^{(1-s)\{(\mathbf{W}+\mathbf{U}) \odot (\mathbf{W}+\mathbf{U})\}} - e^{(1-s)(\mathbf{W} \odot \mathbf{W})} \right] (2\mathbf{W} \odot \mathbf{U}) e^{s(\mathbf{W} \odot \mathbf{W})} \, ds \right\|_F \\
&\le \int_0^1 (3M e^{M^2} \|\mathbf{U}\|_F)(2M \|\mathbf{U}\|_F) e^{M^2} \, ds \\
&\le 6M^2 e^{2M^2} \|\mathbf{U}\|_F^2.
\end{aligned}$$

Thus

$$\left\| e^{(\mathbf{W}+\mathbf{U})\odot(\mathbf{W}+\mathbf{U})} - e^{\mathbf{W}\odot\mathbf{W}} - \int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \right\|_F$$
$$\leq (e^{M^2} + 6M^2e^{2M^2})\|\mathbf{U}\|_F^2.$$

Consequently, the product $\mathbf{W} \odot (e^{\mathbf{W}\odot\mathbf{W}})^\top$ has the relation

$$\begin{aligned}
&\left\| (\mathbf{W}+\mathbf{U})_{\mathcal{A}_j,j} \odot \left[\left(e^{(\mathbf{W}+\mathbf{U})\odot(\mathbf{W}+\mathbf{U})} \right)^\top \right]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} \right. \\
&\quad \left. - \mathbf{U}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \right)^\top \right]_{\mathcal{A}_j,j} \right\|_2 \\
&\leq \left\| \mathbf{U}_{\mathcal{A}_j,j} \odot \left[\left(e^{(\mathbf{W}+\mathbf{U})\odot(\mathbf{W}+\mathbf{U})} - e^{\mathbf{W}\odot\mathbf{W}} \right)^\top \right]_{\mathcal{A}_j,j} \right\|_2 \\
&\quad + \left\| \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(e^{(\mathbf{W}+\mathbf{U})\odot(\mathbf{W}+\mathbf{U})} - e^{\mathbf{W}\odot\mathbf{W}} - \int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \right)^\top \right]_{\mathcal{A}_j,j} \right\|_2 \\
&\leq \|\mathbf{U}\|_F \left\| e^{(\mathbf{W}+\mathbf{U})\odot(\mathbf{W}+\mathbf{U})} - e^{\mathbf{W}\odot\mathbf{W}} \right\|_F + \|\mathbf{W}\|_F(e^{M^2} + 6M^2e^{2M^2})\|\mathbf{U}\|_F^2 \\
&\leq \|\mathbf{U}\|_F \int_0^1 e^{(1-s)M^2}(3M\|\mathbf{U}\|_F)e^{sM^2}\,ds + M(e^{M^2} + 6M^2e^{2M^2})\|\mathbf{U}\|_F^2 \\
&\leq \left[3Me^{M^2} + Me^{M^2} + 6M^3e^{2M^2} \right] \|\mathbf{U}\|_F^2 \\
&\leq \left[4Me^{M^2} + 6M^3e^{2M^2} \right] \|\mathbf{U}\|_F^2.
\end{aligned}$$

The preceding estimates prove that $\mathcal{F}$ is Fréchet differentiable in the active coordinates, with active derivative in direction $\mathbf{U}_{\mathcal{A}^c} = \mathbf{0}$ equal to

$$\left(\begin{aligned}
&\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^d \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r}) \right]^\top \mathbf{U}_{\mathcal{A}_r,r} \\
&\quad + \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{U}^\top(\mathbf{I}-\mathbf{W})^{-\top} \right]_{\mathcal{A}_j,j} + 2\lambda_2 \left\{ \mathbf{U}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} \right. \\
&\qquad \left. + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \right)^\top \right]_{\mathcal{A}_j,j} \right\}
\end{aligned} \right)_{j=1}^d .$$

The same finite-dimensional estimates applied to repeated increments show real analyticity of $\mathcal{F}$ on the active chart. Since the point is transition-free, the active derivative is an isomorphism. Equivalently, the following coercivity quantity is strictly positive as

$$\begin{aligned}
0 < \inf_{\mathbf{U}_{\mathcal{A}^c}=\mathbf{0},\ \|\mathbf{U}\|_F=1} \sum_{j=1}^d \left\| \right. &\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^d \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r}) \right]^\top \mathbf{U}_{\mathcal{A}_r,r} \\
&+ \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{U}^\top(\mathbf{I}-\mathbf{W})^{-\top} \right]_{\mathcal{A}_j,j} + 2\lambda_2 \left\{ \mathbf{U}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} \right. \\
&\left. \left. + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W}\odot\mathbf{U})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \right)^\top \right]_{\mathcal{A}_j,j} \right\} \right\|_2^2 .
\end{aligned}$$

Thus the analytic implicit-function theorem on finite-dimensional Banach spaces applies to $\mathcal{F}(\mathbf{W}_{\mathcal{A}}, \lambda_1, \lambda_2) = \mathbf{0}$ [Hytönen et al., 2016]. Hence, locally on a transition-free chart, the stationary branch is analytic in the relevant scalar parameter, and its difference quotients converge in the active coordinate space.

We now derive the λ_1-flow through finite differences. Fix $\lambda_2 = \lambda_2^0$. For $t \neq 0$, define $\mathbf{W}^t = \mathbf{W}(\lambda_1 + t)$, $\Delta_t \mathbf{W} = \dfrac{\mathbf{W}^t - \mathbf{W}}{t}$, and $\Delta_t \mathbf{W}_{\mathcal{A}^c} = \mathbf{0}$. The active stationary equations at $\lambda_1 + t$ and λ_1 are $\mathcal{F}((\mathbf{W}^t)_{\mathcal{A}}, \lambda_1 + t, \lambda_2^0) = \mathbf{0}$ and $\mathcal{F}(\mathbf{W}_{\mathcal{A}}, \lambda_1, \lambda_2^0) = \mathbf{0}$. Subtracting and dividing by t gives, for every $j = 1, \dots, d$,

$$\begin{aligned}
\mathbf{0} = \frac{1}{t}\left[\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}^t - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}\right] \\
+ \frac{[(\mathbf{I} - \mathbf{W}^t)^{-\top}]_{\mathcal{A}_j,j} - [(\mathbf{I} - \mathbf{W})^{-\top}]_{\mathcal{A}_j,j}}{t} \\
+ 2\lambda_2^0 \frac{\mathbf{W}_{\mathcal{A}_j,j}^t \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}}{t} + \operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j}).
\end{aligned}$$

The residual quotient has the exact finite expansion

$$\begin{aligned}
&\frac{1}{t}\left[\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}^t - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}\right] \\
&= \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}(\mathbf{W}_{:,j}^t - \mathbf{W}_{:,j})}{t\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} + d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})\frac{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2}{t\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} \\
&= \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j}\Delta_t \mathbf{W}_{\mathcal{A}_j,j}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}\frac{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}{t}.
\end{aligned}$$

The residual denominator difference is

$$\begin{aligned}
&\frac{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}{t} \\
&= \frac{1}{t}\sum_{r=1}^d \left[\|\mathbf{X}_{:,r} - \mathbf{X}\mathbf{W}_{:,r}^t\|_2^2 - \|\mathbf{X}_{:,r} - \mathbf{X}\mathbf{W}_{:,r}\|_2^2\right] \\
&= \frac{1}{t}\sum_{r=1}^d \left[\|\mathbf{X}_{:,r} - \mathbf{X}\mathbf{W}_{:,r} - t\mathbf{X}\Delta_t\mathbf{W}_{:,r}\|_2^2 - \|\mathbf{X}_{:,r} - \mathbf{X}\mathbf{W}_{:,r}\|_2^2\right] \\
&= -2\sum_{r=1}^d (\mathbf{X}_{:,r} - \mathbf{X}\mathbf{W}_{:,r})^\top \mathbf{X}\Delta_t\mathbf{W}_{:,r} + t\sum_{r=1}^d \|\mathbf{X}\Delta_t\mathbf{W}_{:,r}\|_2^2 \\
&= 2\sum_{r=1}^d (\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})^\top \mathbf{X}_{:,\mathcal{A}_r}\Delta_t\mathbf{W}_{\mathcal{A}_r,r} + t\sum_{r=1}^d \|\mathbf{X}_{:,\mathcal{A}_r}\Delta_t\mathbf{W}_{\mathcal{A}_r,r}\|_2^2 \\
&= 2\sum_{r=1}^d \left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})\right]^\top \Delta_t\mathbf{W}_{\mathcal{A}_r,r} + t\sum_{r=1}^d \|\mathbf{X}_{:,\mathcal{A}_r}\Delta_t\mathbf{W}_{\mathcal{A}_r,r}\|_2^2.
\end{aligned}$$

The inverse-transpose quotient is exactly

$$\begin{aligned}
\frac{(\mathbf{I} - \mathbf{W}^t)^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top}}{t} &= \frac{1}{t}(\mathbf{I} - \mathbf{W}^t)^{-\top}\left[(\mathbf{I} - \mathbf{W})^\top - (\mathbf{I} - \mathbf{W}^t)^\top\right](\mathbf{I} - \mathbf{W})^{-\top} \\
&= (\mathbf{I} - \mathbf{W}^t)^{-\top}(\Delta_t\mathbf{W})^\top(\mathbf{I} - \mathbf{W})^{-\top}.
\end{aligned}$$

The exponential quotient is

$$\begin{aligned}
\frac{e^{\mathbf{W}^t\odot\mathbf{W}^t} - e^{\mathbf{W}\odot\mathbf{W}}}{t} &= \int_0^1 e^{(1-s)(\mathbf{W}^t\odot\mathbf{W}^t)}\frac{\mathbf{W}^t\odot\mathbf{W}^t - \mathbf{W}\odot\mathbf{W}}{t}e^{s(\mathbf{W}\odot\mathbf{W})}\,ds \\
&= \int_0^1 e^{(1-s)(\mathbf{W}^t\odot\mathbf{W}^t)}\left[2\mathbf{W}\odot\Delta_t\mathbf{W} + t\,\Delta_t\mathbf{W}\odot\Delta_t\mathbf{W}\right]e^{s(\mathbf{W}\odot\mathbf{W})}\,ds,
\end{aligned}$$

and therefore

$$\begin{aligned}
&\frac{\mathbf{W}^t_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}}{t} \\
&= \Delta_t \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} \\
&\quad + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}^t \odot \mathbf{W}^t)} \left[2\mathbf{W} \odot \Delta_t \mathbf{W} + t\, \Delta_t \mathbf{W} \odot \Delta_t \mathbf{W} \right] e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right)^\top \right]_{\mathcal{A}_j,j}.
\end{aligned}$$

By the analytic implicit-function theorem, $\Delta_t \mathbf{W} \to \mathbf{V}^{(1)}$, $\mathbf{V}^{(1)}_{\mathcal{A}^c} = \mathbf{0}$. After possibly shrinking the t-neighborhood, enlarge M by the finite amount $\sup_{0<|t|\le\rho} \left\| \frac{\mathbf{W}(\lambda_1+t) - \mathbf{W}(\lambda_1)}{t} \right\|_F$, so the same symbol M also satisfies $\|\Delta_t \mathbf{W}\|_F \le M$ for $0 < |t| \le \rho$. Then the second-order denominator term vanishes because

$$\begin{aligned}
\left| t \sum_{r=1}^d \|\mathbf{X}_{:,\mathcal{A}_r} \Delta_t \mathbf{W}_{\mathcal{A}_r,r}\|_2^2 \right| &\le |t| \sum_{r=1}^d \|\mathbf{X}_{:,\mathcal{A}_r}\|_{\mathrm{op}}^2 \|\Delta_t \mathbf{W}_{\mathcal{A}_r,r}\|_2^2 \\
&\le |t| \|\mathbf{X}\|_{\mathrm{op}}^2 \sum_{r=1}^d \|\Delta_t \mathbf{W}_{\mathcal{A}_r,r}\|_2^2 \\
&\le |t| M^2 \|\Delta_t \mathbf{W}\|_F^2 \\
&\le |t| M^4 \to 0.
\end{aligned}$$

The exponential integral converges because

$$\begin{aligned}
&\left\| \int_0^1 e^{(1-s)(\mathbf{W}^t \odot \mathbf{W}^t)} \left[2\mathbf{W} \odot \Delta_t \mathbf{W} + t\, \Delta_t \mathbf{W} \odot \Delta_t \mathbf{W} \right] e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right. \\
&\qquad \left. - \int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})} (2\mathbf{W} \odot \mathbf{V}^{(1)}) e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right\|_F \\
&\le \int_0^1 \left\| e^{(1-s)(\mathbf{W}^t \odot \mathbf{W}^t)} - e^{(1-s)(\mathbf{W}\odot\mathbf{W})} \right\|_{\mathrm{op}} \|2\mathbf{W} \odot \Delta_t \mathbf{W} + t\, \Delta_t \mathbf{W} \odot \Delta_t \mathbf{W}\|_F\, e^{sM^2}\, ds \\
&\quad + \int_0^1 e^{(1-s)M^2} \left\| 2\mathbf{W} \odot (\Delta_t \mathbf{W} - \mathbf{V}^{(1)}) + t\, \Delta_t \mathbf{W} \odot \Delta_t \mathbf{W} \right\|_F e^{sM^2}\, ds \\
&\le \int_0^1 e^{(1-s)M^2} \|\mathbf{W}^t \odot \mathbf{W}^t - \mathbf{W} \odot \mathbf{W}\|_F e^{(1-s)M^2} \left[2M \|\Delta_t \mathbf{W}\|_F + |t| \|\Delta_t \mathbf{W}\|_F^2 \right] e^{sM^2}\, ds \\
&\quad + e^{M^2} \left[2M \|\Delta_t \mathbf{W} - \mathbf{V}^{(1)}\|_F + |t| \|\Delta_t \mathbf{W}\|_F^2 \right] \\
&\le e^{2M^2} \|\mathbf{W}^t - \mathbf{W}\|_F (\|\mathbf{W}^t\|_F + \|\mathbf{W}\|_F)(2M^2 + |t|M^2) + e^{M^2} \left[2M \|\Delta_t \mathbf{W} - \mathbf{V}^{(1)}\|_F + |t| M^2 \right] \\
&\le e^{2M^2} (|t| \|\Delta_t \mathbf{W}\|_F)(2M)(3M^2) + e^{M^2} \left[2M \|\Delta_t \mathbf{W} - \mathbf{V}^{(1)}\|_F + |t| M^2 \right] \\
&\le 6|t| M^4 e^{2M^2} + 2M e^{M^2} \|\Delta_t \mathbf{W} - \mathbf{V}^{(1)}\|_F + |t| M^2 e^{M^2} \to 0.
\end{aligned}$$

Taking $t \to 0$ in the finite-difference stationary identity yields

$$\begin{aligned}
&\frac{d\, \mathbf{X}^\top_{:,\mathcal{A}_j} \mathbf{X}_{:,\mathcal{A}_j} \mathbf{V}^{(1)}_{\mathcal{A}_j,j}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\, \mathbf{X}^\top_{:,\mathcal{A}_j} (\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^d \left[\mathbf{X}^\top_{:,\mathcal{A}_r} (\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r}) \right]^\top \mathbf{V}^{(1)}_{\mathcal{A}_r,r} \\
&\quad + \left[(\mathbf{I} - \mathbf{W})^{-\top} (\mathbf{V}^{(1)})^\top (\mathbf{I} - \mathbf{W})^{-\top} \right]_{\mathcal{A}_j,j} \\
&\quad + 2\lambda_2^0 \left\{ \mathbf{V}^{(1)}_{\mathcal{A}_j,j} \odot \left[(e^{\mathbf{W}\odot\mathbf{W}})^\top \right]_{\mathcal{A}_j,j} + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})} \left(2\mathbf{W} \odot \mathbf{V}^{(1)} \right) e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right)^\top \right]_{\mathcal{A}_j,j} \right\} \\
&= -\operatorname{sign}(\mathbf{W}_{\mathcal{A}_j,j}), \qquad j = 1, \ldots, d.
\end{aligned}$$

Since $\mathbf{V}^{(1)}$ is the limit of the active difference quotient, $\dfrac{d\mathbf{W}_{\mathcal{A}_j,j}}{d\lambda_1} = \mathbf{V}^{(1)}_{\mathcal{A}_j,j}$ and $\dfrac{d\mathbf{W}_{\mathcal{A}^c}}{d\lambda_1} = \mathbf{0}$. This proves the λ_1-matrix flow.

The λ_2-flow is obtained by the same finite-difference mechanism, with the additional explicit forcing caused by the varying multiplier of $h(\mathbf{W})$. Fix $\lambda_1 = \lambda_1^0$, and for $t \neq 0$ define $\mathbf{W}^t = \mathbf{W}(\lambda_2 + t)$, $\Delta_t \mathbf{W} = \dfrac{\mathbf{W}^t - \mathbf{W}}{t}$, and $\Delta_t \mathbf{W}_{\mathcal{A}^c} = \mathbf{0}$. The stationary equations $\mathcal{F}((\mathbf{W}^t)_{\mathcal{A}}, \lambda_1^0, \lambda_2 + t) = \mathbf{0}$ and $\mathcal{F}(\mathbf{W}_{\mathcal{A}}, \lambda_1^0, \lambda_2) = \mathbf{0}$ give

$$\mathbf{0} = \frac{1}{t}\left[\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}^t - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}\right] + \frac{[(\mathbf{I} - \mathbf{W}^t)^{-\top}]_{\mathcal{A}_j,j} - [(\mathbf{I} - \mathbf{W})^{-\top}]_{\mathcal{A}_j,j}}{t}$$
$$+ 2\lambda_2 \frac{\mathbf{W}^t_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}}{t} + 2\mathbf{W}^t_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j}.$$

The first three finite quotients are exactly

$$\frac{1}{t}\left[\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j}^t - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2}\right]$$
$$= \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \Delta_t \mathbf{W}_{\mathcal{A}_j,j}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2} - \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}_{:,j} - \mathbf{X}_{:,j})}{\|\mathbf{X} - \mathbf{X}\mathbf{W}^t\|_F^2 \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} \cdot$$
$$\left\{2\sum_{r=1}^{d}\left[\mathbf{X}_{:,\mathcal{A}_r}^\top(\mathbf{X}\mathbf{W}_{:,r} - \mathbf{X}_{:,r})\right]^\top \Delta_t \mathbf{W}_{\mathcal{A}_r,r} + t\sum_{r=1}^{d}\|\mathbf{X}_{:,\mathcal{A}_r}\Delta_t \mathbf{W}_{\mathcal{A}_r,r}\|_2^2\right\},$$

$\dfrac{(\mathbf{I} - \mathbf{W}^t)^{-\top} - (\mathbf{I} - \mathbf{W})^{-\top}}{t} = (\mathbf{I} - \mathbf{W}^t)^{-\top}(\Delta_t \mathbf{W})^\top(\mathbf{I} - \mathbf{W})^{-\top}$, and

$$\frac{\mathbf{W}^t_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} - \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}}{t}$$
$$= \Delta_t \mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j}$$
$$+ \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}^t \odot \mathbf{W}^t)}\left[2\mathbf{W} \odot \Delta_t \mathbf{W} + t\,\Delta_t \mathbf{W} \odot \Delta_t \mathbf{W}\right] e^{s(\mathbf{W}\odot\mathbf{W})}\,ds\right)^\top\right]_{\mathcal{A}_j,j}.$$

Again the analytic implicit-function theorem gives $\Delta_t \mathbf{W} \to \mathbf{V}^{(2)}$, $\mathbf{V}^{(2)}_{\mathcal{A}^c} = \mathbf{0}$. After enlarging M by $\sup_{0<|t|\le\rho}\left\|\dfrac{\mathbf{W}(\lambda_2 + t) - \mathbf{W}(\lambda_2)}{t}\right\|_F$, the same estimates used for the λ_1-flow give $t\sum_{r=1}^{d}\|\mathbf{X}_{:,\mathcal{A}_r}\Delta_t \mathbf{W}_{\mathcal{A}_r,r}\|_2^2 \to 0$, $(\mathbf{I} - \mathbf{W}^t)^{-\top}(\Delta_t \mathbf{W})^\top(\mathbf{I} - \mathbf{W})^{-\top} \to (\mathbf{I} - \mathbf{W})^{-\top}(\mathbf{V}^{(2)})^\top(\mathbf{I} - \mathbf{W})^{-\top}$, and

$$\int_0^1 e^{(1-s)(\mathbf{W}^t \odot \mathbf{W}^t)}\left[2\mathbf{W} \odot \Delta_t \mathbf{W} + t\,\Delta_t \mathbf{W} \odot \Delta_t \mathbf{W}\right] e^{s(\mathbf{W}\odot\mathbf{W})}\,ds$$
$$\longrightarrow \int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})}(2\mathbf{W} \odot \mathbf{V}^{(2)})e^{s(\mathbf{W}\odot\mathbf{W})}\,ds.$$

The explicit forcing term satisfies

$$\left\|2\mathbf{W}^t_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}^t \odot \mathbf{W}^t})^\top]_{\mathcal{A}_j,j} - 2\mathbf{W}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j}\right\|_2$$
$$\le 2\|\mathbf{W}^t - \mathbf{W}\|_F \|e^{\mathbf{W}^t \odot \mathbf{W}^t}\|_{\mathrm{op}} + 2\|\mathbf{W}\|_F \|e^{\mathbf{W}^t \odot \mathbf{W}^t} - e^{\mathbf{W}\odot\mathbf{W}}\|_F$$
$$\le 2|t|\|\Delta_t \mathbf{W}\|_F e^{M^2} + 2M\int_0^1 e^{(1-s)M^2}\|\mathbf{W}^t \odot \mathbf{W}^t - \mathbf{W} \odot \mathbf{W}\|_F e^{sM^2}\,ds$$
$$\le 2|t|Me^{M^2} + 2Me^{M^2}\|\mathbf{W}^t - \mathbf{W}\|_F(\|\mathbf{W}^t\|_F + \|\mathbf{W}\|_F)$$
$$\le 2|t|Me^{M^2} + 4|t|M^3e^{M^2} \to 0.$$

Therefore, taking $t \to 0$ in the λ_2-difference equation gives

$$\begin{aligned}
&\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \mathbf{V}^{(2)}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^{d} \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{X}\mathbf{W}_{:,r}-\mathbf{X}_{:,r})\right]^\top \mathbf{V}^{(2)}_{\mathcal{A}_r,r} \\
&\quad + \left[(\mathbf{I}-\mathbf{W})^{-\top}(\mathbf{V}^{(2)})^\top (\mathbf{I}-\mathbf{W})^{-\top}\right]_{\mathcal{A}_j,j} \\
&\quad + 2\lambda_2 \left\{ \mathbf{V}^{(2)}_{\mathcal{A}_j,j} \odot \left[(e^{\mathbf{W}\odot\mathbf{W}})^\top\right]_{\mathcal{A}_j,j} + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})} \big(2\mathbf{W}\odot \mathbf{V}^{(2)}\big) e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right)^\top \right]_{\mathcal{A}_j,j} \right\} \\
&= -2\mathbf{W}_{\mathcal{A}_j,j} \odot \left[(e^{\mathbf{W}\odot\mathbf{W}})^\top\right]_{\mathcal{A}_j,j}, \qquad j=1,\dots,d.
\end{aligned}$$

Since $\mathbf{V}^{(2)}$ is the limit of the active λ_2-difference quotient, $\dfrac{d\mathbf{W}_{\mathcal{A}_j,j}}{d\lambda_2} = \mathbf{V}^{(2)}_{\mathcal{A}_j,j}$ and $\dfrac{d\mathbf{W}_{\mathcal{A}^c}}{d\lambda_2} = \mathbf{0}$. This proves the λ_2-matrix flow.

Finally, we show that the velocity in each displayed system is unique. If two active velocities solve the same linear velocity system at the same $(\mathbf{W}, \lambda_1, \lambda_2)$, their difference $\mathbf{U}$, with $\mathbf{U}_{\mathcal{A}^c} = \mathbf{0}$, satisfies the homogeneous active linearized system. The coercivity condition gives

$$\begin{aligned}
0 = \sum_{j=1}^{d} \Bigg\| &\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \mathbf{U}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^{d} \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{X}\mathbf{W}_{:,r}-\mathbf{X}_{:,r})\right]^\top \mathbf{U}_{\mathcal{A}_r,r} \Bigg\|_2^2 \\
&+ \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{U}^\top (\mathbf{I}-\mathbf{W})^{-\top}\right]_{\mathcal{A}_j,j} + 2\lambda_2 \Bigg\{ \mathbf{U}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} \\
&\qquad + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})} (2\mathbf{W}\odot \mathbf{U}) e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right)^\top \right]_{\mathcal{A}_j,j} \Bigg\}
\end{aligned}$$

$$\geq \left[\inf_{\mathbf{Z}_{\mathcal{A}^c}=\mathbf{0},\ \|\mathbf{Z}\|_F=1} \sum_{j=1}^{d} \Bigg\| \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top \mathbf{X}_{:,\mathcal{A}_j} \mathbf{Z}_{\mathcal{A}_j,j}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} - \frac{2d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4} \sum_{r=1}^{d} \left[\mathbf{X}_{:,\mathcal{A}_r}^\top (\mathbf{X}\mathbf{W}_{:,r}-\mathbf{X}_{:,r})\right]^\top \mathbf{Z}_{\mathcal{A}_r,r} \Bigg\|_2^2 \right] \|\mathbf{U}\|_F^2.$$

$$\begin{aligned}
&+ \left[(\mathbf{I}-\mathbf{W})^{-\top}\mathbf{Z}^\top (\mathbf{I}-\mathbf{W})^{-\top}\right]_{\mathcal{A}_j,j} + 2\lambda_2 \Bigg\{ \mathbf{Z}_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}\odot\mathbf{W}})^\top]_{\mathcal{A}_j,j} \\
&\qquad + \mathbf{W}_{\mathcal{A}_j,j} \odot \left[\left(\int_0^1 e^{(1-s)(\mathbf{W}\odot\mathbf{W})} (2\mathbf{W}\odot \mathbf{Z}) e^{s(\mathbf{W}\odot\mathbf{W})}\, ds \right)^\top \right]_{\mathcal{A}_j,j} \Bigg\}
\end{aligned}$$

The bracketed infimum is strictly positive on the transition-free regular chart, hence $\|\mathbf{U}\|_F = 0$. Thus the active velocity is unique.

The same finite-difference construction also proves claimed exactness. For the λ_1-flow, the active residual is constant along the transition-free branch; since it is zero at the initial stationary point, for every λ_1 in the same transition-free interval,

$$\begin{aligned}
&\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}(\lambda_1)_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1)\|_F^2} + [(\mathbf{I}-\mathbf{W}(\lambda_1))^{-\top}]_{\mathcal{A}_j,j} + 2\lambda_2^0 \mathbf{W}(\lambda_1)_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}(\lambda_1)\odot\mathbf{W}(\lambda_1)})^\top]_{\mathcal{A}_j,j} + \lambda_1 \operatorname{sign}(\mathbf{W}(\lambda_1)_{\mathcal{A}_j,j}) \\
&= \frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top (\mathbf{X}\mathbf{W}(\lambda_1^0)_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_1^0)\|_F^2} + [(\mathbf{I}-\mathbf{W}(\lambda_1^0))^{-\top}]_{\mathcal{A}_j,j} + 2\lambda_2^0 \mathbf{W}(\lambda_1^0)_{\mathcal{A}_j,j} \odot [(e^{\mathbf{W}(\lambda_1^0)\odot\mathbf{W}(\lambda_1^0)})^\top]_{\mathcal{A}_j,j} + \lambda_1^0 \operatorname{sign}(\mathbf{W}(\lambda_1^0)_{\mathcal{A}_j,j}) = \mathbf{0}.
\end{aligned}$$

For the λ_2-flow,

$$\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}(\lambda_2)_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_2)\|_F^2}+[(\mathbf{I}-\mathbf{W}(\lambda_2))^{-\top}]_{\mathcal{A}_j,j}+2\lambda_2\mathbf{W}(\lambda_2)_{\mathcal{A}_j,j}\odot[(e^{\mathbf{W}(\lambda_2)\odot\mathbf{W}(\lambda_2)})^\top]_{\mathcal{A}_j,j}+\lambda_1^0\operatorname{sign}(\mathbf{W}(\lambda_2)_{\mathcal{A}_j,j})$$
$$=\frac{d\,\mathbf{X}_{:,\mathcal{A}_j}^\top(\mathbf{X}\mathbf{W}(\lambda_2^0)_{:,j}-\mathbf{X}_{:,j})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}(\lambda_2^0)\|_F^2}+[(\mathbf{I}-\mathbf{W}(\lambda_2^0))^{-\top}]_{\mathcal{A}_j,j}+2\lambda_2^0\mathbf{W}(\lambda_2^0)_{\mathcal{A}_j,j}\odot[(e^{\mathbf{W}(\lambda_2^0)\odot\mathbf{W}(\lambda_2^0)})^\top]_{\mathcal{A}_j,j}+\lambda_1^0\operatorname{sign}(\mathbf{W}(\lambda_2^0)_{\mathcal{A}_j,j})=\mathbf{0}.$$

The inactive coordinates remain zero because $\mathbf{W}_{\mathcal{A}^c}\equiv\mathbf{0}$, and the inactive subgradient inequalities remain valid until the next transition point by the transition-free hypothesis. Hence the two displayed systems characterize the exact stationary matrix flows of (3) on the transition-free interval. □

J Proof of Theorem 4.4

Proof sketch. We spell out the invariant-regime argument, since the detailed algebra follows the same spirit as Appendix F. Fix one coordinate path and write $\theta=\lambda_q$, $q\in\{1,2\}$, with the other regularization parameter held fixed. On a non-transition segment, the active support $\mathcal{A}$ and the active signs are fixed, $\mathbf{W}_{\mathcal{A}^c}\equiv\mathbf{0}$, the GOLEM domain conditions $\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2>0$ and $\det(\mathbf{I}-\mathbf{W})\neq 0$ hold, and the active linearization $\mathcal{L}_{\lambda_2}$ is provably nonsingular.

Recall that Theorems 4.2 and 4.3 gives a locally unique analytic continuation of the stationary branch in the same active chart. For any θ strictly before the minimum displayed in Theorem 4.4, none of the boundary events in Appendix A.1 can occur: no active coefficient reaches zero, every inactive subgradient inequality remains strict (with threshold λ_1 along the λ_1-path and threshold λ_1^0 along the λ_2-path), the log-domain does not degenerate, and $\mathcal{L}_{\lambda_2}$ has no nontrivial kernel. Therefore the same fixed-regime KKT system continues to hold, so the branch cannot encounter a transition before that minimum. Conversely, once the minimum is attained, the branch reaches one of the only possible obstructions to staying in the same analytic regime: support/sign loss, inactive activation, smooth-domain failure, loss of invertibility of the velocity system, or, on the λ_1-path, the endpoint $\lambda_{\max}$. These alternatives are precisely $\lambda_{1,\mathrm{drop}}$, $\lambda_{1,\mathrm{enter}}$, $\lambda_{1,\mathrm{dom}}$, $\lambda_{1,\mathrm{turn}}$, $\lambda_{\max}$, and analogously $\lambda_{2,\mathrm{drop}}$, $\lambda_{2,\mathrm{enter}}$, $\lambda_{2,\mathrm{dom}}$, $\lambda_{2,\mathrm{turn}}$ for the λ_2-path. Unlike NOTEARS, no additional topological-order case is needed because $h(\mathbf{W})$ enters GOLEM only as a smooth penalty. The remaining root and margin calculations follow the same strategy as the proof of Appendix F, with only minor modifications, and are omitted. □

K Proof of Theorem 5.1

Proof. Let $\rho(\cdot)$ denote spectral radius, and interpret the equality constraint $h^s_{\mathrm{ldet}}(\mathbf{W})=0$ on the admissible DAGMA domain, that is, on the set of matrices for which $s\mathbf{I}-\mathbf{W}\odot\mathbf{W}$ is in the required M-matrix log-determinant domain, equivalently $\rho(\mathbf{W}\odot\mathbf{W})<s$. For notational compactness inside the proof only, write

$$\mathsf{F}_\lambda(\mathbf{W}):=\frac{1}{2n}\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2+\lambda\|\mathbf{W}\|_1,\qquad \Delta_\lambda(\mathbf{W}):=\mathsf{F}_\lambda(\mathbf{W})-\mathsf{F}_\lambda(\mathbf{0}).$$

First, $\mathbf{0}$ is feasible because

$$h^s_{\mathrm{ldet}}(\mathbf{0})=-\log\det(s\mathbf{I}-\mathbf{0}\odot\mathbf{0})+d\log s=-\log\det(s\mathbf{I})+d\log s=0. \tag{32}$$

Let $\mathbf{W}$ be any feasible point. Since $\rho(\mathbf{W}\odot\mathbf{W})<s$, the logarithm admits the finite-dimensional holomorphic functional calculus expansion. For every $\mathbf{A}$ with $\rho(\mathbf{A})<1$,

$$-\log\det(\mathbf{I}-\mathbf{A})=-\operatorname{tr}\log(\mathbf{I}-\mathbf{A})=\sum_{k=1}^{\infty}\frac{1}{k}\operatorname{tr}(\mathbf{A}^k).$$

Applying this with $\mathbf{A}=s^{-1}(\mathbf{W}\odot\mathbf{W})$, we obtain the exact closed-walk expansion

$$\begin{aligned}0=h^s_{\mathrm{ldet}}(\mathbf{W})&=-\log\left[s^d\det\left(\mathbf{I}-\frac{1}{s}\mathbf{W}\odot\mathbf{W}\right)\right]+d\log s\\&=-d\log s-\log\det\left(\mathbf{I}-\frac{1}{s}\mathbf{W}\odot\mathbf{W}\right)+d\log s\\&=-\operatorname{tr}\log\left(\mathbf{I}-\frac{1}{s}\mathbf{W}\odot\mathbf{W}\right)\end{aligned}$$

$$= \sum_{k=1}^{\infty} \frac{1}{k} \operatorname{tr}\left[\left(\frac{1}{s}\mathbf{W} \odot \mathbf{W}\right)^k\right]$$
$$= \sum_{k=1}^{\infty} \frac{1}{ks^k} \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \cdots \sum_{i_k=1}^{d} \mathbf{W}_{i_1 i_2}^2 \mathbf{W}_{i_2 i_3}^2 \cdots \mathbf{W}_{i_{k-1} i_k}^2 \mathbf{W}_{i_k i_1}^2,$$

which is a series of nonnegative closed-walk monomials. In particular, the $k = 1$ contribution gives $0 = \sum_{i=1}^{d} \mathbf{W}_{ii}^2$. Thus the DAGMA equality constraint implies the exact off-diagonal reduction

$$\|\mathbf{W}\|_1 = \sum_{i=1}^{d} \sum_{j=1}^{d} |\mathbf{W}_{ij}| = \sum_{\substack{1 \le i,j \le d \\ i \ne j}} |\mathbf{W}_{ij}|.$$

We next record the precise finite-dimensional dual certificate induced by $\lambda_{\max}$. Since $\mathbf{X}^\top \mathbf{X}/n$ is symmetric and every feasible $\mathbf{W}$ has zero diagonal,

$$\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\, \mathbf{W}\right) = \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)_{ij} \mathbf{W}_{ji} = \sum_{\substack{1 \le i,j \le d \\ i \ne j}} \left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)_{ij} \mathbf{W}_{ij}.$$

Consequently, the off-diagonal ℓ_1/ℓ_∞ polar identity yields

$$\sup_{\substack{\mathbf{W} \in \mathbb{R}^{d \times d} \\ \mathbf{W}_{ii} = 0,\ \|\mathbf{W}\|_1 \le 1}} \operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\, \mathbf{W}\right) = \sup_{\substack{\mathbf{W} \in \mathbb{R}^{d \times d} \\ \mathbf{W}_{ii} = 0,\ \sum_{i \ne j} |\mathbf{W}_{ij}| \le 1}} \sum_{\substack{1 \le i,j \le d \\ i \ne j}} \left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)_{ij} \mathbf{W}_{ij}$$
$$= \max_{\substack{1 \le i,j \le d \\ i \ne j}} \left|\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)_{ij}\right|$$
$$= \left\|\frac{1}{n}\mathbf{X}^\top \mathbf{X} - \operatorname{diag}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)\right\|_{\max} = \lambda_{\max}.$$

So for every feasible $\mathbf{W}$,

$$\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\, \mathbf{W}\right) \le \left|\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\, \mathbf{W}\right)\right|$$
$$\le \sum_{\substack{1 \le i,j \le d \\ i \ne j}} \left|\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right)_{ij}\right| |\mathbf{W}_{ij}|$$
$$\le \lambda_{\max} \sum_{\substack{1 \le i,j \le d \\ i \ne j}} |\mathbf{W}_{ij}| = \lambda_{\max} \|\mathbf{W}\|_1.$$

Now expand the constrained objective gap. Since $h^s_{\mathrm{ldet}}(\mathbf{W}) = h^s_{\mathrm{ldet}}(\mathbf{0}) = 0$ for feasible $\mathbf{W}$, we may insert the equality-constraint residual into the comparison without changing the value

$$\begin{aligned}
\Delta_\lambda(\mathbf{W}) &= \mathsf{F}_\lambda(\mathbf{W}) - \mathsf{F}_\lambda(\mathbf{0}) \\
&= \left\{\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1 + h^s_{\mathrm{ldet}}(\mathbf{W})\right\} - \left\{\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{0}\|_F^2 + \lambda\|\mathbf{0}\|_1 + h^s_{\mathrm{ldet}}(\mathbf{0})\right\} \\
&= \frac{1}{2n}\operatorname{tr}\big[(\mathbf{X} - \mathbf{X}\mathbf{W})^\top (\mathbf{X} - \mathbf{X}\mathbf{W})\big] - \frac{1}{2n}\operatorname{tr}(\mathbf{X}^\top \mathbf{X}) + \lambda\|\mathbf{W}\|_1 + h^s_{\mathrm{ldet}}(\mathbf{W}) \\
&= \frac{1}{2n}\operatorname{tr}\big[\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}\mathbf{W} - \mathbf{W}^\top \mathbf{X}^\top \mathbf{X} + \mathbf{W}^\top \mathbf{X}^\top \mathbf{X}\mathbf{W}\big] - \frac{1}{2n}\operatorname{tr}(\mathbf{X}^\top \mathbf{X}) + \lambda\|\mathbf{W}\|_1 + h^s_{\mathrm{ldet}}(\mathbf{W}) \\
&= \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 - \operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{X}\, \mathbf{W}\right) + \lambda\|\mathbf{W}\|_1 + h^s_{\mathrm{ldet}}(\mathbf{W}).
\end{aligned} \tag{33}$$

Using both the off-diagonal reduction forced by $h^s_{\mathrm{ldet}}(\mathbf{W}) = 0$ and the closed-walk representation of $h^s_{\mathrm{ldet}}(\mathbf{W})$, (33) becomes

$$\Delta_\lambda(\mathbf{W}) = \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 - \sum_{\substack{1\le i,j\le d\\ i\neq j}} \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij} \mathbf{W}_{ij} + \lambda \sum_{\substack{1\le i,j\le d\\ i\neq j}} |\mathbf{W}_{ij}| + \sum_{k=1}^{\infty} \frac{1}{ks^k} \sum_{i_1=1}^{d}\sum_{i_2=1}^{d}\cdots\sum_{i_k=1}^{d} \mathbf{W}^2_{i_1i_2}\mathbf{W}^2_{i_2i_3}\cdots\mathbf{W}^2_{i_{k-1}i_k}\mathbf{W}^2_{i_ki_1}.$$

For each off-diagonal entry, decompose the linear-ℓ_1 contribution as

$$\lambda|\mathbf{W}_{ij}| - \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\mathbf{W}_{ij} = (\lambda-\lambda_{\max})|\mathbf{W}_{ij}| + \left[\lambda_{\max} - \left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right|\right]|\mathbf{W}_{ij}| + \left[\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right||\mathbf{W}_{ij}| - \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\mathbf{W}_{ij}\right].$$

Substitution gives the exact nonnegative decomposition

$$\begin{aligned}\Delta_\lambda(\mathbf{W}) = \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 &+ (\lambda-\lambda_{\max})\sum_{\substack{1\le i,j\le d\\ i\neq j}}|\mathbf{W}_{ij}| + \sum_{\substack{1\le i,j\le d\\ i\neq j}}\left[\lambda_{\max} - \left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right|\right]|\mathbf{W}_{ij}|\\ &+ \sum_{\substack{1\le i,j\le d\\ i\neq j}}\left[\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right||\mathbf{W}_{ij}| - \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\mathbf{W}_{ij}\right]\\ &+ \sum_{k=1}^{\infty}\frac{1}{ks^k}\sum_{i_1=1}^{d}\sum_{i_2=1}^{d}\cdots\sum_{i_k=1}^{d}\mathbf{W}^2_{i_1i_2}\mathbf{W}^2_{i_2i_3}\cdots\mathbf{W}^2_{i_{k-1}i_k}\mathbf{W}^2_{i_ki_1}.\end{aligned}$$

When $\lambda \ge \lambda_{\max}$, every term on the right-hand side is nonnegative as $\frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 \ge 0$ and $(\lambda-\lambda_{\max})\sum_{i\neq j}|\mathbf{W}_{ij}| \ge 0$. For every $i \neq j$, $\lambda_{\max} - \left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right| \ge 0$ and $\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right||\mathbf{W}_{ij}| - \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\mathbf{W}_{ij} \ge 0$, and the final closed-walk series is a sum of nonnegative monomials. Therefore, for every feasible $\mathbf{W}$,

$$\Delta_\lambda(\mathbf{W}) \ge \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 + (\lambda-\lambda_{\max})\|\mathbf{W}\|_1 \ge 0.$$

Since $\mathbf{0}$ is feasible by (32), this proves that $\mathbf{0}$ is a global minimizer for every $\lambda \ge \lambda_{\max}$.

We now formally prove uniqueness through a global error-bound argument. First suppose $\lambda > \lambda_{\max}$. For every feasible $\mathbf{W}$, $\Delta_\lambda(\mathbf{W}) \ge (\lambda-\lambda_{\max})\|\mathbf{W}\|_1$. Hence if $\mathbf{W} \neq \mathbf{0}$, then $\|\mathbf{W}\|_1 > 0$, and therefore $\Delta_\lambda(\mathbf{W}) > 0$. Thus no nonzero feasible point can attain the objective value at $\mathbf{0}$, proving uniqueness when $\lambda > \lambda_{\max}$. More strongly, if $\{\mathbf{W}_m\}_{m\ge1}$ is any feasible minimizing sequence satisfying $\Delta_\lambda(\mathbf{W}_m) \to 0$, then

$$0 \le \|\mathbf{W}_m\|_1 \le \frac{\Delta_\lambda(\mathbf{W}_m)}{\lambda-\lambda_{\max}} \longrightarrow 0,$$

and hence every such sequence converges to $\mathbf{0}$ in the entrywise ℓ_1 norm. Equivalently, for every $\varepsilon > 0$, choosing $\delta = (\lambda-\lambda_{\max})\varepsilon$ gives the deterministic implication that, for every feasible $\mathbf{W}$, $\Delta_\lambda(\mathbf{W}) < \delta$ implies $\|\mathbf{W}\|_1 < \varepsilon$. It remains to handle the boundary case $\lambda = \lambda_{\max}$ with $\lambda_{\max} > 0$. Let $\{\mathbf{W}_m\}_{m\ge1}$ be any feasible sequence satisfying $\Delta_{\lambda_{\max}}(\mathbf{W}_m) \to 0$. The nonnegative decomposition above gives $0 \le \frac{1}{2n}\|\mathbf{X}\mathbf{W}_m\|_F^2 \le \Delta_{\lambda_{\max}}(\mathbf{W}_m)$, and hence $\|\mathbf{X}\mathbf{W}_m\|_F \le \sqrt{2n\,\Delta_{\lambda_{\max}}(\mathbf{W}_m)} \longrightarrow$

0. On the other hand, by (8) the exact gap identity at $\lambda = \lambda_{\max}$ yields

$$\begin{aligned}\lambda_{\max}\|\mathbf{W}_m\|_1 &= \Delta_{\lambda_{\max}}(\mathbf{W}_m) - \frac{1}{2n}\|\mathbf{X}\mathbf{W}_m\|_F^2 + \operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_m\right)\\ &\le \Delta_{\lambda_{\max}}(\mathbf{W}_m) + \left|\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}_m\right)\right|\\ &\le \Delta_{\lambda_{\max}}(\mathbf{W}_m) + \frac{1}{n}\|\mathbf{X}\|_F\|\mathbf{X}\mathbf{W}_m\|_F\\ &\le \Delta_{\lambda_{\max}}(\mathbf{W}_m) + \frac{1}{n}\|\mathbf{X}\|_F\sqrt{2n\,\Delta_{\lambda_{\max}}(\mathbf{W}_m)}.\end{aligned}$$

Since $\lambda_{\max} > 0$, division by $\lambda_{\max}$ gives the quantitative stability estimate

$$0 \le \|\mathbf{W}_m\|_1 \le \frac{\Delta_{\lambda_{\max}}(\mathbf{W}_m)}{\lambda_{\max}} + \frac{\|\mathbf{X}\|_F}{\lambda_{\max}}\sqrt{\frac{2\,\Delta_{\lambda_{\max}}(\mathbf{W}_m)}{n}} \longrightarrow 0.$$

Equivalently, for every $\varepsilon > 0$, choose $\delta > 0$ small enough that $\dfrac{\delta}{\lambda_{\max}} + \dfrac{\|\mathbf{X}\|_F}{\lambda_{\max}}\sqrt{\dfrac{2\delta}{n}} < \varepsilon$. Then every feasible $\mathbf{W}$ satisfying $\Delta_{\lambda_{\max}}(\mathbf{W}) < \delta$ also satisfies $\|\mathbf{W}\|_1 < \varepsilon$. In particular, if $\mathbf{W}$ is a feasible global minimizer at $\lambda = \lambda_{\max} > 0$, then $\Delta_{\lambda_{\max}}(\mathbf{W}) = 0$, and the preceding deterministic error bound gives $\|\mathbf{W}\|_1 = 0$, hence $\mathbf{W} = \mathbf{0}$. This proves uniqueness in the boundary case $\lambda = \lambda_{\max} > 0$. Combining the strict and boundary cases proves that the global minimizer is unique whenever $\lambda > \lambda_{\max}$ or $\lambda_{\max} > 0$.

Finally, the value $\lambda_{\max}$ is sharp as a zero-solution threshold. Let $\mathbf{e}_i$ denote the i-th standard basis vector in $\mathbb{R}^d$, and define $\mathbf{E}_{ij} := \mathbf{e}_i\mathbf{e}_j^\top$, the matrix with a single unit entry at position (i,j) and zeros elsewhere. Suppose $0 \le \lambda < \lambda_{\max}$. Then $\lambda_{\max} > 0$, so there exist $i \ne j$ such that $\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right| = \lambda_{\max}$. For $t > 0$, define

$$\mathbf{W} = t\,\operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\mathbf{E}_{ij}.$$

Since $i \ne j$,

$$\mathbf{W}\odot\mathbf{W} = t^2\mathbf{E}_{ij}, \quad (\mathbf{W}\odot\mathbf{W})^2 = t^4\mathbf{E}_{ij}^2 = t^4\mathbf{e}_i(\mathbf{e}_j^\top\mathbf{e}_i)\mathbf{e}_j^\top = \mathbf{0}, \quad \rho(\mathbf{W}\odot\mathbf{W}) = 0 < s.$$

Moreover,

$$\begin{aligned}h^s_{\mathrm{ldet}}(\mathbf{W}) &= -\log\det(s\mathbf{I} - t^2\mathbf{E}_{ij}) + d\log s = -\log\det\left[s\left(\mathbf{I} - \frac{t^2}{s}\mathbf{E}_{ij}\right)\right] + d\log s\\ &= -\log\left[s^d\det\left(\mathbf{I} - \frac{t^2}{s}\mathbf{e}_i\mathbf{e}_j^\top\right)\right] + d\log s = 0,\end{aligned}$$

so this one-edge perturbation is feasible. For this feasible path, $\|\mathbf{W}\|_1 = t$, and

$$\begin{aligned}\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) &= t\,\operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{E}_{ij}\right)\\ &= t\,\operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ji}\\ &= t\left|\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right| = t\lambda_{\max},\end{aligned}$$

where symmetry of $\mathbf{X}^\top\mathbf{X}/n$ is used in the penultimate equality, and

$$\begin{aligned}\frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 &= \frac{1}{2n}\left\|t\,\operatorname{sign}\left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ij}\right)\mathbf{X}\mathbf{E}_{ij}\right\|_F^2\\ &= \frac{t^2}{2n}\|\mathbf{X}\mathbf{e}_i\mathbf{e}_j^\top\|_F^2\\ &= \frac{t^2}{2n}\|\mathbf{X}\mathbf{e}_i\|_2^2\|\mathbf{e}_j\|_2^2 = \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Delta_\lambda(\mathbf{W}) &= \frac{1}{2n}\|\mathbf{X}\mathbf{W}\|_F^2 - \operatorname{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,\mathbf{W}\right) + \lambda\|\mathbf{W}\|_1 \\ &= \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii} - t\lambda_{\max} + \lambda t \\ &= \frac{t^2}{2}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii} - t(\lambda_{\max} - \lambda).\end{aligned}$$

Because $\lambda_{\max} - \lambda > 0$ and $(\mathbf{X}^\top\mathbf{X}/n)_{ii} \geq 0$, the last expression is strictly negative for all sufficiently small $t > 0$, for example for any

$$0 < t < \begin{cases} \dfrac{2(\lambda_{\max} - \lambda)}{\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii}}, & \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii} > 0, \\ +\infty, & \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{ii} = 0. \end{cases}$$

Thus, for every $0 \leq \lambda < \lambda_{\max}$, there exists a feasible nonzero $\mathbf{W}$ with $\mathsf{F}_\lambda(\mathbf{W}) < \mathsf{F}_\lambda(\mathbf{0})$. Hence no smaller value than $\lambda_{\max}$ can guarantee the zero solution, and the stated threshold is exact. □

L Proof of Theorem 6.1

We first prove the finite transition-point bounds for NOTEARS and DAGMA in Part I below, and then establish the finite transition-point bound for GOLEM in Part II (Appendix L.2).

L.1 Part I

Proof. Let $\mathfrak{S}_d$ denote the symmetric group on $\{1, \ldots, d\}$, let $\operatorname{adj}(M)$ denote the adjugate matrix of a square matrix M, i.e., $M^{-1} = \operatorname{adj}(M)/\det(M)$ whenever M is nonsingular, and for an index set $\mathcal{B} \subseteq \{1, \ldots, d\}$ let $\mathbf{X}_{\cdot\mathcal{B}}$ denote the submatrix of $\mathbf{X}$ with columns indexed by $\mathcal{B}$. For a sign vector $\sigma \in \{-1, 1\}^{\mathcal{B}}$ and a subset $\mathcal{C} \subseteq \mathcal{B}$, let $\sigma_{\mathcal{C}}$ denote the restriction of σ to coordinates in $\mathcal{C}$.

If $d = 1$, then $h(\mathbf{W}) = \operatorname{tr}(e^{\mathbf{W}\odot\mathbf{W}}) - 1 = e^{\mathbf{W}_{11}^2} - 1$, whence $h(\mathbf{W}) = 0$ implies $\mathbf{W}_{11} = 0$. The stationary graph is the single analytic ray $\{(0, \lambda) : 0 < \lambda < +\infty\}$ and has no transition point. The asserted right-hand side in Theorem 6.1 equals

$$1 + 1!\left[\frac{1}{2}\left(\sum_{r=0}^{\min(n,0)} 2^r\binom{0}{r} - 1\right) - 1\right] = 1 + \left[\frac{1}{2}(1 - 1) - 1\right] = 0,$$

so the assertion holds.

Henceforth we assume $d \geq 2$ and define $\lambda_{\max} := \max_{1\leq i<j\leq d}\left|\frac{1}{n}\mathbf{X}_{\cdot i}^\top\mathbf{X}_{\cdot j}\right|$. We first construct the full-measure event on which the whole argument is valid. Let $z = \operatorname{vec}(\mathbf{X}) \in \mathbb{R}^{nd}$. Since z admits a density with respect to Lebesgue measure, for every nonzero polynomial $p : \mathbb{R}^{nd} \to \mathbb{R}$, $\mathbb{P}\{p(z) = 0\} = \int_{\{u\in\mathbb{R}^{nd}:p(u)=0\}} f(u)\,du = 0$, because the zero set of a nonzero polynomial has Lebesgue measure zero. Since there are only finitely many subsets $\mathcal{B} \subseteq \{1, \ldots, d\}$, indices $q \notin \mathcal{B}$, signs $\sigma \in \{-1, 1\}^{\mathcal{B}}$, and signs $\tau \in \{-1, 1\}$, the following event has probability one:

$$\begin{aligned}
&\det\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right) > 0, && 1 \leq |\mathcal{B}| \leq n, \\
&\sum_{\substack{\mathcal{C}\subseteq\mathcal{B} \\ |\mathcal{C}|=n+1}} \det\left(\begin{bmatrix}\mathbf{X}_{\cdot\mathcal{C}} \\ \sigma_{\mathcal{C}}^\top\end{bmatrix}\right)^2 > 0, && n+1 \leq |\mathcal{B}| \leq d, \quad \sigma \in \{-1, 1\}^{\mathcal{B}}, \\
&\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\operatorname{adj}\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\sigma \neq \tau\det\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right), && 1 \leq |\mathcal{B}| \leq n, \quad q \notin \mathcal{B}, \quad \sigma \in \{-1, 1\}^{\mathcal{B}}, \quad \tau \in \{-1, 1\}, \\
&\mathbf{X}_{\cdot i}^\top\mathbf{X}_{\cdot j} \neq 0, && 1 \leq i < j \leq d.
\end{aligned} \tag{34}$$

Equivalently, each equality excluded above is the zero set of a nonzero polynomial. For the first family, the polynomial $\det(\mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}})$ is not identically zero because one may set the columns in $\mathcal{B}$ equal to distinct coordinate vectors. For the second family, choose $\mathcal{C} = \{c_1, \dots, c_{n+1}\} \subseteq \mathcal{B}$ and set

$$\mathbf{X}_{\cdot c_1} = e_1, \quad \mathbf{X}_{\cdot c_2} = e_2, \quad \cdots, \quad \mathbf{X}_{\cdot c_n} = e_n, \quad \mathbf{X}_{\cdot c_{n+1}} = 0,$$

where $e_1, \dots, e_n$ are the canonical basis vectors of $\mathbb{R}^n$; then

$$\det \begin{bmatrix} e_1 & e_2 & \cdots & e_n & 0 \\ \sigma_{c_1} & \sigma_{c_2} & \cdots & \sigma_{c_n} & \sigma_{c_{n+1}} \end{bmatrix} = (-1)^n \sigma_{c_{n+1}} \neq 0.$$

For the third family, after taking $\det(\mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}}) \neq 0$ and $\mathbf{X}_{\cdot q} = 0$, the left-hand side is zero and the right-hand side is nonzero. For the fourth family, set $\mathbf{X}_{\cdot i} = \mathbf{X}_{\cdot j} = e_1$. Thus every excluded equality is a genuine algebraic degeneracy, and hence all displayed relations hold simultaneously almost surely. All subsequent arguments are restricted to this full-measure event. Based on (34), we can derive the exact stratification induced by the NOTEARS constraint. For $\mathbf{A} := \mathbf{W} \odot \mathbf{W}$, the exponential series (5) gives

$$h(\mathbf{W}) = \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{i_0=1}^{d} \sum_{i_1=1}^{d} \cdots \sum_{i_{m-1}=1}^{d} \mathbf{W}_{i_0 i_1}^2 \mathbf{W}_{i_1 i_2}^2 \cdots \mathbf{W}_{i_{m-1} i_0}^2,$$

where every summand is nonnegative. Therefore $h(\mathbf{W}) = 0$ holds exactly when $\mathbf{W}_{i_0 i_1}^2 \mathbf{W}_{i_1 i_2}^2 \cdots \mathbf{W}_{i_{m-1} i_0}^2 = 0$ for every $m \geq 1$ and every $(i_0, \dots, i_{m-1})$. This is equivalent to the existence of $\pi \in \mathfrak{S}_d$ satisfying $\mathbf{W}_{\pi(b),\pi(a)} = 0$ for all $1 \leq a \leq b \leq d$. For each $\pi \in \mathfrak{S}_d$, define the closed triangular stratum

$$\mathcal{D}_\pi := \left\{ \mathbf{W} \in \mathbb{R}^{d \times d} : \mathbf{W}_{\pi(b),\pi(a)} = 0,\ 1 \leq a \leq b \leq d \right\}.$$

Then the preceding equivalence is the exact identity $\{\mathbf{W} : h(\mathbf{W}) = 0\} = \bigcup_{\pi \in \mathfrak{S}_d} \mathcal{D}_\pi$. Moreover, on $\mathcal{D}_\pi$, the matrix $\mathbf{A} := \mathbf{W} \odot \mathbf{W}$ is strictly upper triangular after the permutation π, and hence $\mathbf{A}^m$ is also strictly upper triangular for $1 \leq m \leq d-1$, while $\mathbf{A}^m = 0$ for $m \geq d$. Consequently, for every $\mathbf{W} \in \mathcal{D}_\pi$, $\operatorname{tr}(\mathbf{A}^m) = 0$ for $m \geq 1$, and $h(\mathbf{W}) = 0$. The first derivative of the acyclicity function is also identically zero on the acyclic set. Indeed, $\dfrac{\partial h(\mathbf{W})}{\partial \mathbf{W}_{ij}} = 2\mathbf{W}_{ij} \left(e^{\mathbf{W} \odot \mathbf{W}}\right)_{ji}$, and if $\mathbf{W}_{ij} = 0$ the derivative is zero, whereas if $\mathbf{W}_{ij} \neq 0$, acyclicity excludes every directed path from j to i, so $\left(e^{\mathbf{W} \odot \mathbf{W}}\right)_{ji} = \displaystyle\sum_{m=0}^{d-1} \frac{((\mathbf{W} \odot \mathbf{W})^m)_{ji}}{m!} = 0$. Thus the analytic equality $h(\mathbf{W}) = 0$ contributes through the exact finite union of linear topological strata $\mathcal{D}_\pi$.

For $\lambda > 0$, write the objective as $\Phi_\lambda(\mathbf{W}) := \dfrac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_1$. On $\mathcal{D}_\pi$, we have

$$\Phi_\lambda(\mathbf{W}) = \sum_{b=1}^{d} \left[\frac{1}{2n} \left\| \mathbf{X}_{\cdot,\pi(b)} - \sum_{a=1}^{b-1} \mathbf{X}_{\cdot,\pi(a)} \mathbf{W}_{\pi(a),\pi(b)} \right\|_2^2 + \lambda \sum_{a=1}^{b-1} \left| \mathbf{W}_{\pi(a),\pi(b)} \right| \right],$$

which is obtained by expanding the Frobenius norm columnwise as

$$\begin{aligned} \frac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 &= \frac{1}{2n} \sum_{j=1}^{d} \left\| \mathbf{X}_{\cdot j} - \sum_{i=1}^{d} \mathbf{X}_{\cdot i} \mathbf{W}_{ij} \right\|_2^2 \\ &= \frac{1}{2n} \sum_{b=1}^{d} \left\| \mathbf{X}_{\cdot,\pi(b)} - \sum_{a=1}^{b-1} \mathbf{X}_{\cdot,\pi(a)} \mathbf{W}_{\pi(a),\pi(b)} - \sum_{a=b}^{d} \mathbf{X}_{\cdot,\pi(a)} \mathbf{W}_{\pi(a),\pi(b)} \right\|_2^2 \\ &= \frac{1}{2n} \sum_{b=1}^{d} \left\| \mathbf{X}_{\cdot,\pi(b)} - \sum_{a=1}^{b-1} \mathbf{X}_{\cdot,\pi(a)} \mathbf{W}_{\pi(a),\pi(b)} \right\|_2^2. \end{aligned} \tag{35}$$

The last equality in (35) uses $\mathbf{W}_{\pi(a),\pi(b)} = 0$ for $a \geq b$ on $\mathcal{D}_\pi$; similarly, $\|\mathbf{W}\|_1 = \sum_{j=1}^{d}\sum_{i=1}^{d}|\mathbf{W}_{ij}| = \sum_{b=1}^{d}\sum_{a=1}^{b-1}|\mathbf{W}_{\pi(a),\pi(b)}|$. Hence, after fixing π, the constrained stationary equations separate into d nodewise Lasso-type stationary systems. Fix $\pi \in \mathfrak{S}_d$ and $k \in \{0,\dots,d-1\}$. The node in position $k+1$ has the possible parent set $\{\pi(1),\dots,\pi(k)\}$. For a coefficient vector $u = (u_1,\dots,u_k)^\top \in \mathbb{R}^k$, define the node-wise criterion

$$\varphi_{\pi,k,\lambda}(u) := \frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}u_a\right\|_2^2 + \lambda\sum_{a=1}^{k}|u_a|.$$

The KKT conditions for u are

$$0 \in \frac{1}{n}\mathbf{X}_{\cdot,\pi(a)}^\top\left(\sum_{\ell=1}^{k}\mathbf{X}_{\cdot,\pi(\ell)}u_\ell - \mathbf{X}_{\cdot,\pi(k+1)}\right) + \lambda\,\partial|u_a|, \qquad a = 1,\dots,k.$$

Let $\mathcal{B} \subseteq \{\pi(1),\dots,\pi(k)\}$ and $\sigma \in \{-1,1\}^{\mathcal{B}}$. If the active parent set is $\mathcal{B}$ and the active sign vector is σ, then the active coordinates must satisfy

$$\begin{aligned}0 &= \frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\left(\mathbf{X}_{\cdot\mathcal{B}}\mathbf{W}_{\mathcal{B},\pi(k+1)} - \mathbf{X}_{\cdot,\pi(k+1)}\right) + \lambda\sigma \\ &= \left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\mathbf{W}_{\mathcal{B},\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} + \lambda\sigma.\end{aligned}$$

If $|\mathcal{B}| \leq n$, then

$$\det\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right) = n^{-|\mathcal{B}|}\det\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right) > 0,$$

and therefore the unique affine candidate in this signed regime is

$$\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda) = \left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma\right), \qquad \mathbf{W}_{q,\pi(k+1)}(\lambda) = 0 \quad (q \notin \mathcal{B}).$$

The relative interior of this regime is the set of $\lambda > 0$ satisfying

$$\begin{aligned}&\sigma_i e_i^\top\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma\right) > 0, \qquad i \in \mathcal{B}, \\ &-\lambda < \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma\right) < \lambda, \\ &q \in \{\pi(1),\dots,\pi(k)\}\setminus\mathcal{B},\end{aligned} \tag{36}$$

where e_i denotes the coordinate vector indexed by i in $\mathbb{R}^{|\mathcal{B}|}$, and each condition is affine in λ. Hence (36) tells

$$\begin{aligned}&\sigma_i e_i^\top\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma\right) \\ &= \sigma_i e_i^\top\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma_i e_i^\top\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\sigma,\end{aligned}$$

and, for each inactive q,

$$\begin{aligned}&\frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma\right) \\ &= \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} + \lambda\frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\sigma \\ &= \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left[n\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\right]\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} + \lambda\frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left[n\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\right]\sigma \\ &= \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \frac{1}{n}\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} + \lambda\mathbf{X}_{\cdot q}^\top\mathbf{X}_{\cdot\mathcal{B}}\left(\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\sigma.\end{aligned}$$

Consequently the inactive equality

$$\frac{1}{n}\mathbf{X}_{\cdot,q}^{\top}\mathbf{X}_{\cdot,\pi(k+1)}-\frac{1}{n}\mathbf{X}_{\cdot,q}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot,\pi(k+1)}-\lambda\sigma\right)=\tau\lambda,\qquad\tau\in\{-1,1\},$$

can hold for all λ in a nonempty open interval only if the coefficient of λ is τ, namely only if $\mathbf{X}_{\cdot,q}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}\sigma=\tau$. Using

$$\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)^{-1}=\frac{\operatorname{adj}\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)}{\det\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)},$$

this is equivalent to

$$\mathbf{X}_{\cdot,q}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\operatorname{adj}\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right)\sigma=\tau\det\left(\mathbf{X}_{\cdot\mathcal{B}}^{\top}\mathbf{X}_{\cdot\mathcal{B}}\right),$$

which is excluded on the full-measure event. Hence no inactive subgradient equality is identically tight along a nontrivial regular branch.

We next prove that an active set of cardinality larger than n cannot occur. Suppose $|\mathcal{B}|\geq n+1$ and the active sign vector is $\sigma\in\{-1,1\}^{\mathcal{B}}$. The active KKT equation gives

$$\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^{\top}\left(\mathbf{X}_{\cdot,\pi(k+1)}-\mathbf{X}_{\cdot\mathcal{B}}\mathbf{W}_{\mathcal{B},\pi(k+1)}\right)=\lambda\sigma.$$

Equivalently,

$$\begin{aligned}\sigma^{\top}&=\frac{1}{n\lambda}\left(\mathbf{X}_{\cdot,\pi(k+1)}-\mathbf{X}_{\cdot\mathcal{B}}\mathbf{W}_{\mathcal{B},\pi(k+1)}\right)^{\top}\mathbf{X}_{\cdot\mathcal{B}}\\&=\left[\frac{1}{n\lambda}\left(\mathbf{X}_{\cdot,\pi(k+1)}-\mathbf{X}_{\cdot\mathcal{B}}\mathbf{W}_{\mathcal{B},\pi(k+1)}\right)\right]^{\top}\mathbf{X}_{\cdot\mathcal{B}}.\end{aligned}$$

Therefore $\sigma^{\top}$ belongs to the row span of $\mathbf{X}_{\cdot\mathcal{B}}$, and hence $\operatorname{rank}\begin{bmatrix}\mathbf{X}_{\cdot\mathcal{B}}\\\sigma^{\top}\end{bmatrix}\leq n$. But the second full-measure condition gives $\operatorname{rank}\begin{bmatrix}\mathbf{X}_{\cdot\mathcal{B}}\\\sigma^{\top}\end{bmatrix}=n+1$, since

$$\sum_{\substack{\mathcal{C}\subseteq\mathcal{B}\\|\mathcal{C}|=n+1}}\det\left(\begin{bmatrix}\mathbf{X}_{\cdot\mathcal{C}}\\\sigma_{\mathcal{C}}^{\top}\end{bmatrix}\right)^{2}>0.$$

This contradiction proves $|\mathcal{B}|\leq\min(n,k)$ for every active parent set appearing in every nodewise stationary system. We also need to verify uniqueness of each nodewise Lasso-type solution on the same event. Let $u,v\in\mathbb{R}^{k}$ both minimize $\varphi_{\pi,k,\lambda}$. For $t\in(0,1)$, convexity yields

$$\begin{aligned}&\varphi_{\pi,k,\lambda}(tu+(1-t)v)\\&=\frac{1}{2n}\left\|t\left(\mathbf{X}_{\cdot,\pi(k+1)}-\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}u_a\right)+(1-t)\left(\mathbf{X}_{\cdot,\pi(k+1)}-\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}v_a\right)\right\|_2^2+\lambda\sum_{a=1}^{k}|tu_a+(1-t)v_a|\\&\leq\frac{t}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)}-\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}u_a\right\|_2^2+\frac{1-t}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)}-\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}v_a\right\|_2^2\\&\qquad+\lambda t\sum_{a=1}^{k}|u_a|+\lambda(1-t)\sum_{a=1}^{k}|v_a|\\&=t\varphi_{\pi,k,\lambda}(u)+(1-t)\varphi_{\pi,k,\lambda}(v).\end{aligned}$$

Since both u and v are minimizers, equality holds throughout. Equality in the squared-norm convexity step gives $\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}u_a=\sum_{a=1}^{k}\mathbf{X}_{\cdot,\pi(a)}v_a$, and equality in every triangle inequality gives $|tu_a+(1-t)v_a|=t|u_a|+(1-t)|v_a|$, $a=1,\dots,k$. Let $\mathcal{B}:=\{\pi(a):u_a\neq0\vee v_a\neq0\}$. For each $i\in\mathcal{B}$, the last equality forces u_i and v_i to have a common sign whenever both are nonzero; define $\sigma_i\in\{-1,1\}$

as this common sign, using the nonzero one if exactly one of u_i, v_i is nonzero. For all $t \in (0,1)$ outside at most finitely many values, the active set of $tu + (1-t)v$ is $\mathcal{B}$ with sign vector σ. The KKT equations for $tu + (1-t)v$ give

$$\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top \left(\mathbf{X}_{\cdot,\pi(k+1)} - \mathbf{X}_{\cdot\mathcal{B}}\{tu_\mathcal{B} + (1-t)v_\mathcal{B}\}\right) = \lambda\sigma. \tag{37}$$

Because the fitted values are equal, $\mathbf{X}_{\cdot\mathcal{B}}u_\mathcal{B} = \mathbf{X}_{\cdot\mathcal{B}}v_\mathcal{B}$, the residual in the preceding (37) is independent of t, and we can solve

$$\sigma^\top = \frac{1}{n\lambda}\left(\mathbf{X}_{\cdot,\pi(k+1)} - \mathbf{X}_{\cdot\mathcal{B}}u_\mathcal{B}\right)^\top \mathbf{X}_{\cdot\mathcal{B}}.$$

If $|\mathcal{B}| \geq n+1$, then $\operatorname{rank}\begin{bmatrix}\mathbf{X}_{\cdot\mathcal{B}} \\ \sigma^\top\end{bmatrix} \leq n$, contradicting the full-measure rank condition. If $|\mathcal{B}| \leq n$, then $\mathbf{X}_{\cdot\mathcal{B}}(u_\mathcal{B} - v_\mathcal{B}) = 0$ and

$$\begin{aligned} 0 &= (u_\mathcal{B} - v_\mathcal{B})^\top \mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}}(u_\mathcal{B} - v_\mathcal{B}) \\ &\geq \lambda_{\min}\left(\mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}}\right)\|u_\mathcal{B} - v_\mathcal{B}\|_2^2, \end{aligned}$$

where $\lambda_{\min}\left(\mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}}\right) > 0$ by the first full-measure condition. Hence $u_\mathcal{B} = v_\mathcal{B}$, and therefore $u = v$. Thus the nodewise lasso solution is unique for every $\lambda > 0$, every π, and every k.

The unique nodewise Lasso-type solution is continuous in λ. Let $\lambda_m \to \lambda > 0$ and let $u(\lambda_m)$ be the unique minimizer of $\varphi_{\pi,k,\lambda_m}$. For all sufficiently large m, $\lambda_m \geq \lambda/2$, and comparison with the zero vector gives

$$\begin{aligned} \frac{\lambda}{2}\|u(\lambda_m)\|_1 &\leq \lambda_m \|u(\lambda_m)\|_1 \\ &\leq \frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)}\right\|_2^2 - \frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^k \mathbf{X}_{\cdot,\pi(a)}u_a(\lambda_m)\right\|_2^2 \\ &\leq \frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)}\right\|_2^2. \end{aligned}$$

Thus

$$\|u(\lambda_m)\|_1 \leq \frac{1}{n\lambda}\left\|\mathbf{X}_{\cdot,\pi(k+1)}\right\|_2^2, \qquad \|u(\lambda_m)\|_2 \leq \|u(\lambda_m)\|_1 \leq \frac{1}{n\lambda}\left\|\mathbf{X}_{\cdot,\pi(k+1)}\right\|_2^2.$$

Every subsequence has a further convergent subsequence. If $u(\lambda_{m_j}) \to \bar{u}$, then for every $z \in \mathbb{R}^k$,

$$\begin{aligned} &\frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^k \mathbf{X}_{\cdot,\pi(a)}\bar{u}_a\right\|_2^2 + \lambda\sum_{a=1}^k |\bar{u}_a| \\ &= \lim_{j\to\infty}\left[\frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^k \mathbf{X}_{\cdot,\pi(a)}u_a(\lambda_{m_j})\right\|_2^2 + \lambda_{m_j}\sum_{a=1}^k |u_a(\lambda_{m_j})|\right] \\ &\leq \lim_{j\to\infty}\left[\frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^k \mathbf{X}_{\cdot,\pi(a)}z_a\right\|_2^2 + \lambda_{m_j}\sum_{a=1}^k |z_a|\right] \\ &= \frac{1}{2n}\left\|\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^k \mathbf{X}_{\cdot,\pi(a)}z_a\right\|_2^2 + \lambda\sum_{a=1}^k |z_a|. \end{aligned}$$

Therefore $\bar{u}$ is the unique minimizer at λ, so $\bar{u} = u(\lambda)$. Since every subsequential limit equals $u(\lambda)$, the whole sequence converges.

For a fixed node with k possible predecessors, the number of possible signed active regimes is $\sum_{r=0}^{\min(n,k)} 2^r \binom{k}{r}$, because the active parent set has size $r \leq \min(n,k)$, has $\binom{k}{r}$ choices, and has 2^r possible signs. Each signed regime is valid on an interval, since it is the intersection of finitely

many strict affine inequalities in the scalar variable λ. The interval cannot split into two disjoint intervals, because an intersection of affine half-lines in $\mathbb{R}$ is convex. Moreover, note that for any nonempty active set $\mathcal{B}$, two opposite signs σ and $-\sigma$ cannot both occur. If they occurred at positive regularization values λ_1, λ_2, then the active KKT equations would give

$$\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1) = \frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} - \lambda_1\sigma,$$
$$\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) = \frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot,\pi(k+1)} + \lambda_2\sigma.$$

Subtracting,

$$\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\left[\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right] = (\lambda_1+\lambda_2)\sigma.$$

Multiplying by the transpose of $\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)$ gives

$$\begin{aligned}
0 &< \left[\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right]^\top\left(\frac{1}{n}\mathbf{X}_{\cdot\mathcal{B}}^\top\mathbf{X}_{\cdot\mathcal{B}}\right)\left[\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right] \\
&= (\lambda_1+\lambda_2)\sigma^\top\left[\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right] \\
&= (\lambda_1+\lambda_2)\left[\sigma^\top\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2) - \sigma^\top\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right] \\
&= (\lambda_1+\lambda_2)\left[-\left\|\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_2)\right\|_1 - \left\|\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda_1)\right\|_1\right] < 0,
\end{aligned}$$

which makes a contradiction. Therefore the number of nonempty signed active intervals for a node with k possible predecessors is bounded by $1 + \frac{1}{2}\sum_{r=1}^{\min(n,k)} 2^r\binom{k}{r} = \frac{1}{2}\left(\sum_{r=0}^{\min(n,k)} 2^r\binom{k}{r} + 1\right)$, where the leading 1 is the empty active set. By continuity of the unique nodewise path, these intervals form a connected finite chain in $(0, +\infty)$. Hence the number of nodewise transition points is at most $\frac{1}{2}\left(\sum_{r=0}^{\min(n,k)} 2^r\binom{k}{r} + 1\right) - 1 = \frac{1}{2}\left(\sum_{r=0}^{\min(n,k)} 2^r\binom{k}{r} - 1\right)$. For a fixed topological order π, the restricted stationary matrix is assembled column by column from the d unique nodewise Lasso-type paths, with $k = 0, 1, \ldots, d-1$ possible predecessors. Let $\mathcal{T}_\pi$ denote the fixed-order transition set. Its cardinality is bounded by the sum of the nodewise bounds

$$\begin{aligned}
\#\mathcal{T}_\pi &\le \sum_{k=0}^{d-1}\frac{1}{2}\left(\sum_{r=0}^{\min(n,k)} 2^r\binom{k}{r} - 1\right) \\
&= \frac{1}{2}\sum_{k=0}^{d-1}\left(\sum_{r=0}^{\min(n,k)} 2^r\binom{k}{r} - 1\right).
\end{aligned}$$

It remains to remove the repeated counting of the terminal zero transition. At $\mathbf{W} = 0$,

$$\begin{aligned}
\frac{\partial}{\partial\mathbf{W}_{ij}}\left[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2\right]_{\mathbf{W}=0} &= \frac{1}{n}\mathbf{X}_{\cdot i}^\top\left[(\mathbf{X}\mathbf{W})_{\cdot j} - \mathbf{X}_{\cdot j}\right]_{\mathbf{W}=0} \\
&= -\frac{1}{n}\mathbf{X}_{\cdot i}^\top\mathbf{X}_{\cdot j}.
\end{aligned}$$

For a fixed order π, the zero matrix satisfies all inactive KKT inequalities precisely when $\left|\frac{1}{n}\mathbf{X}_{\cdot,\pi(a)}^\top\mathbf{X}_{\cdot,\pi(b)}\right| \le \lambda$, $1 \le a < b \le d$. Since every topological order contains exactly one orientation of every unordered pair,

$$\begin{aligned}
\max_{1\le a<b\le d}\left|\frac{1}{n}\mathbf{X}_{\cdot,\pi(a)}^\top\mathbf{X}_{\cdot,\pi(b)}\right| &= \max_{1\le i<j\le d}\left|\frac{1}{n}\mathbf{X}_{\cdot i}^\top\mathbf{X}_{\cdot j}\right| \\
&= \lambda_{\max}.
\end{aligned}$$

Thus $\mathbf{W} = 0$ is stationary for every fixed order if and only if $\lambda \geq \lambda_{\max}$, and $(\mathbf{0}, \lambda_{\max})$ belongs to every fixed-order transition set. The same calculation also gives $\lambda_{\max} > 0$ on the full-measure event, because at least one off-diagonal inner product is nonzero when $d \geq 2$.

Now let $\mathcal{T}$ denote the transition set of the NOTEARS (or DAGMA) stationary graph in $0 < \lambda < +\infty$, and use $\mathcal{T}_\pi$ for the transition set of the fixed-order path on $\mathcal{D}_\pi$. Since $\{\mathbf{W} : h(\mathbf{W}) = 0\} = \bigcup_{\pi \in \mathfrak{S}_d} \mathcal{D}_\pi$, every NOTEARS (or DAGMA) stationary point belongs to at least one fixed-order stratum. On the relative interior of any such stratum, the stationarity equations are the nodewise Lasso-type KKT equations derived above. A sign transition is an endpoint of one of the nodewise signed-support intervals. A topological transition is a point at which no single frozen topological order continues to represent the same local regular regime; since the feasible set is exactly the finite union of the strata $\mathcal{D}_\pi$, such a point is contained in the boundary of at least one fixed-order stratum, and the boundary event is realized by either an active coefficient equality $\mathbf{W}_{\pi(a),\pi(b)}(\lambda) = 0$ with $a < b$, or an inactive KKT equality

$$\left| \frac{1}{n} \mathbf{X}_{\cdot q}^\top \left(\mathbf{X}_{\cdot,\pi(k+1)} - \sum_{a=1}^{k} \mathbf{X}_{\cdot,\pi(a)} \mathbf{W}_{\pi(a),\pi(k+1)}(\lambda) \right) \right| = \lambda,$$

for some fixed order π, node position $k+1$, and inactive predecessor $q \in \{\pi(1), \ldots, \pi(k)\}$. Both events are endpoints of the fixed-order nodewise intervals already counted. Therefore $\mathcal{T} \subseteq \bigcup_{\pi \in \mathfrak{S}_d} \mathcal{T}_\pi$.

Because each $\mathcal{T}_\pi$ contains the common point $(\mathbf{0}, \lambda_{\max})$,

$$\begin{aligned}
\#\mathcal{T} &\leq \# \left(\bigcup_{\pi \in \mathfrak{S}_d} \mathcal{T}_\pi \right) \\
&\leq 1 + \sum_{\pi \in \mathfrak{S}_d} (\#\mathcal{T}_\pi - 1) \\
&\leq 1 + \sum_{\pi \in \mathfrak{S}_d} \left[\frac{1}{2} \sum_{k=0}^{d-1} \left(\sum_{r=0}^{\min(n,k)} 2^r \binom{k}{r} - 1 \right) - 1 \right] = 1 + d! \left[\frac{1}{2} \sum_{k=0}^{d-1} \left(\sum_{r=0}^{\min(n,k)} 2^r \binom{k}{r} - 1 \right) - 1 \right].
\end{aligned}$$

Finally, on every fixed triple consisting of a topological order π, an active parent set $\mathcal{B}$, and an active sign vector σ, the branch formula

$$\mathbf{W}_{\mathcal{B},\pi(k+1)}(\lambda) = \left(\frac{1}{n} \mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot\mathcal{B}} \right)^{-1} \left(\frac{1}{n} \mathbf{X}_{\cdot\mathcal{B}}^\top \mathbf{X}_{\cdot,\pi(k+1)} - \lambda\sigma \right)$$

is affine in λ, hence real analytic. Since the number of topological orders, active parent sets, and active sign vectors is finite, the NOTEARS (or DAGMA) stationary solution graph is a finite union of real-analytic arcs. The stated finite analytic stratification and the claimed upper bound follow. □

L.2 Part II

Proof. Without loss of generality, let the continuation branches be restricted to the compact part of the GOLEM effective domain on which

$$\|\mathbf{W}\|_F \leq M, \qquad \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 \geq \eta, \qquad |\det(\mathbf{I} - \mathbf{W})| \geq \eta$$

for some small $\eta > 0$. For every $(i, j) \in \{1, \ldots, d\}^2$, we already have

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}_{ij}} \left\{ \frac{d}{2} \log \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 - \log |\det(\mathbf{I} - \mathbf{W})| + \lambda_2 h(\mathbf{W}) \right\} \\
&= \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W} - \mathbf{X})]_{ij}}{\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I} - \mathbf{W})^{-T}]_{ij} + 2\lambda_2 \mathbf{W}_{ij} (e^{\mathbf{W} \odot \mathbf{W}})_{ji}.
\end{aligned}$$

Here

$$d\{-\log |\det(\mathbf{I} - \mathbf{W})|\} = -\operatorname{tr}\big\{(\mathbf{I} - \mathbf{W})^{-1} d(\mathbf{I} - \mathbf{W})\big\} = \operatorname{tr}\big\{(\mathbf{I} - \mathbf{W})^{-1} d\mathbf{W}\big\} = \sum_{i=1}^{d} \sum_{j=1}^{d} [(\mathbf{I} - \mathbf{W})^{-T}]_{ij} \, d\mathbf{W}_{ij}.$$

Therefore the stationary inclusion is

$$0 \in \frac{d[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}+\lambda_1\partial|\mathbf{W}_{ij}|, \qquad (i,j)\in\{1,\dots,d\}^2,$$

and, on a support-sign regime S, s,

$$\begin{aligned}
&\mathbf{W}_{ij}=0 \quad (i,j)\notin S, \qquad s_{ij}\mathbf{W}_{ij}>0 \quad (i,j)\in S,\\
&\frac{d[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}+\lambda_1 s_{ij}=0, \quad (i,j)\in S,\\
&\left|\frac{d[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}\right|\le\lambda_1, \quad (i,j)\notin S,
\end{aligned} \qquad (38)$$

which reduces on inactive coordinates as follows. Since $\mathbf{W}_{ij}=0$, the penalty-gradient term satisfies $2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}=0$, and the remaining condition is

$$\left|\frac{d[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}\right|\le\lambda_1.$$

We show that all analytic terms are uniformly controlled on the compact effective-domain region. To see this, we have

$$\begin{aligned}
\left|\frac{d[\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{XW}\|_F^2}\right| &\le \frac{d\|\mathbf{X}^\top(\mathbf{XW}-\mathbf{X})\|_F}{\|\mathbf{X}-\mathbf{XW}\|_F^2}\\
&\le \frac{d\|\mathbf{X}\|_2(\|\mathbf{XW}\|_F+\|\mathbf{X}\|_F)}{\eta}\le\frac{d\|\mathbf{X}\|_2(M\|\mathbf{X}\|_2+\|\mathbf{X}\|_F)}{\eta},\\
\|(\mathbf{I}-\mathbf{W})^{-T}\|_F &= \frac{\|\operatorname{adj}(\mathbf{I}-\mathbf{W})\|_F}{|\det(\mathbf{I}-\mathbf{W})|}\\
&\le \frac{\left(\sum_{a=1}^d\sum_{b=1}^d\left|\det((\mathbf{I}-\mathbf{W})_{\{1,\dots,d\}\setminus\{a\},\{1,\dots,d\}\setminus\{b\}})\right|^2\right)^{1/2}}{\eta}\\
&\le \frac{\left(\sum_{a=1}^d\sum_{b=1}^d\{(d-1)!\max_{u,v}|(\mathbf{I}-\mathbf{W})_{uv}|^{d-1}\}^2\right)^{1/2}}{\eta}\\
&\le \frac{d(d-1)!\|\mathbf{I}-\mathbf{W}\|_F^{d-1}}{\eta}\\
&\le \frac{d(d-1)!(\|\mathbf{I}\|_F+\|\mathbf{W}\|_F)^{d-1}}{\eta}\le\frac{d(d-1)!(\sqrt{d}+M)^{d-1}}{\eta},
\end{aligned}$$

and

$$\rho(\mathbf{W}\odot\mathbf{W})\le\|\mathbf{W}\odot\mathbf{W}\|_2\le\|\mathbf{W}\odot\mathbf{W}\|_F=\left(\sum_{u=1}^d\sum_{v=1}^d\mathbf{W}_{uv}^4\right)^{1/2}\le\sum_{u=1}^d\sum_{v=1}^d\mathbf{W}_{uv}^2=\|\mathbf{W}\|_F^2\le M^2,$$

$$\|e^{\mathbf{W}\odot\mathbf{W}}\|_2=\left\|\sum_{q=0}^\infty\frac{(\mathbf{W}\odot\mathbf{W})^q}{q!}\right\|_2\le\sum_{q=0}^\infty\frac{\|\mathbf{W}\odot\mathbf{W}\|_2^q}{q!}\le\sum_{q=0}^\infty\frac{\|\mathbf{W}\odot\mathbf{W}\|_F^q}{q!}\le\sum_{q=0}^\infty\frac{M^{2q}}{q!}=e^{M^2},$$

$$|2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}|\le 2|\lambda_2|\,|\mathbf{W}_{ij}|\|e^{\mathbf{W}\odot\mathbf{W}}\|_2\le 2|\lambda_2|\|\mathbf{W}\|_F e^{M^2}\le 2|\lambda_2|Me^{M^2}.$$

For $\|\mathbf{W}+\mathbf{\Delta}\|_F\le M+1$, it follows that

$$\begin{aligned}
&\left\|e^{(\mathbf{W}+\mathbf{\Delta})\odot(\mathbf{W}+\mathbf{\Delta})}-e^{\mathbf{W}\odot\mathbf{W}}\right\|_2\\
&\le\int_0^1\left\|e^{(1-t)(\mathbf{W}\odot\mathbf{W})}\right\|_2\|\mathbf{W}\odot\mathbf{\Delta}+\mathbf{\Delta}\odot\mathbf{W}+\mathbf{\Delta}\odot\mathbf{\Delta}\|_F\left\|e^{t((\mathbf{W}+\mathbf{\Delta})\odot(\mathbf{W}+\mathbf{\Delta}))}\right\|_2 dt\\
&\le\int_0^1 e^{(1-t)M^2}\{\|\mathbf{W}\odot\mathbf{\Delta}\|_F+\|\mathbf{\Delta}\odot\mathbf{W}\|_F+\|\mathbf{\Delta}\odot\mathbf{\Delta}\|_F\}e^{t(M+1)^2}dt\\
&\le e^{M^2+(M+1)^2}\{2\|\mathbf{W}\|_F\|\mathbf{\Delta}\|_F+\|\mathbf{\Delta}\|_F^2\}\le e^{M^2+(M+1)^2}(2M+1)\|\mathbf{\Delta}\|_F,
\end{aligned}$$

and

$$
\begin{aligned}
&\left\| e^{(\mathbf{W}+\mathbf{\Delta})\odot(\mathbf{W}+\mathbf{\Delta})} - e^{\mathbf{W}\odot\mathbf{W}} - \int_0^1 e^{(1-t)(\mathbf{W}\odot\mathbf{W})}(\mathbf{W}\odot\mathbf{\Delta}+\mathbf{\Delta}\odot\mathbf{W})e^{t(\mathbf{W}\odot\mathbf{W})}\,dt \right\|_2 \\
&\le \int_0^1 \|e^{(1-t)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{\Delta}\odot\mathbf{\Delta}\|_F \|e^{t(\mathbf{W}\odot\mathbf{W})}\|_2\,dt \\
&\quad + \int_0^1\int_0^1 \|e^{(1-u)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{W}\odot\mathbf{\Delta}+\mathbf{\Delta}\odot\mathbf{W}\|_F \|e^{(u-v)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{W}\odot\mathbf{\Delta}+\mathbf{\Delta}\odot\mathbf{W}\|_F \|e^{v(\mathbf{W}\odot\mathbf{W})}\|_2\,dv\,du \\
&\quad + 2\int_0^1\int_0^1 \|e^{(1-u)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{W}\odot\mathbf{\Delta}+\mathbf{\Delta}\odot\mathbf{W}\|_F \|e^{(u-v)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{\Delta}\odot\mathbf{\Delta}\|_F \|e^{v(\mathbf{W}\odot\mathbf{W})}\|_2\,dv\,du \\
&\quad + \int_0^1\int_0^1 \|e^{(1-u)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{\Delta}\odot\mathbf{\Delta}\|_F \|e^{(u-v)(\mathbf{W}\odot\mathbf{W})}\|_2 \|\mathbf{\Delta}\odot\mathbf{\Delta}\|_F \|e^{v(\mathbf{W}\odot\mathbf{W})}\|_2\,dv\,du \\
&\le e^{M^2}\|\mathbf{\Delta}\|_F^2 + e^{M^2}(2M\|\mathbf{\Delta}\|_F)^2 + 2e^{M^2}(2M\|\mathbf{\Delta}\|_F)\|\mathbf{\Delta}\|_F^2 + e^{M^2}\|\mathbf{\Delta}\|_F^4 \\
&= e^{M^2}\{(1+4M^2)\|\mathbf{\Delta}\|_F^2 + 4M\|\mathbf{\Delta}\|_F^3 + \|\mathbf{\Delta}\|_F^4\}.
\end{aligned}
$$

Meanwhile, it holds that

$$
\begin{aligned}
&\|(\mathbf{I}-\mathbf{W}-\mathbf{\Delta})^{-T} - (\mathbf{I}-\mathbf{W})^{-T}\|_F \\
&= \|(\mathbf{I}-\mathbf{W}-\mathbf{\Delta})^{-T}\mathbf{\Delta}^T(\mathbf{I}-\mathbf{W})^{-T}\|_F \\
&\le \|(\mathbf{I}-\mathbf{W}-\mathbf{\Delta})^{-T}\|_F \|\mathbf{\Delta}\|_F \|(\mathbf{I}-\mathbf{W})^{-T}\|_F \\
&\le \frac{2d(d-1)!(\sqrt{d}+M+1)^{d-1}}{\eta}\|\mathbf{\Delta}\|_F \frac{d(d-1)!(\sqrt{d}+M)^{d-1}}{\eta} \\
&= \frac{2d^2((d-1)!)^2(\sqrt{d}+M+1)^{d-1}(\sqrt{d}+M)^{d-1}}{\eta^2}\|\mathbf{\Delta}\|_F,
\end{aligned}
$$

and

$$
\begin{aligned}
&\left| \frac{d[\mathbf{X}^\top(\mathbf{X}(\mathbf{W}+\mathbf{\Delta})-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}(\mathbf{W}+\mathbf{\Delta})\|_F^2} - \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} \right| \\
&\le d\frac{\|\mathbf{X}^\top\mathbf{X}\mathbf{\Delta}\|_F}{\|\mathbf{X}-\mathbf{X}(\mathbf{W}+\mathbf{\Delta})\|_F^2} + d\|\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})\|_F \frac{\left|\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2 - \|\mathbf{X}-\mathbf{X}(\mathbf{W}+\mathbf{\Delta})\|_F^2\right|}{\|\mathbf{X}-\mathbf{X}(\mathbf{W}+\mathbf{\Delta})\|_F^2\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} \\
&\le \frac{2d\|\mathbf{X}\|_2^2}{\eta}\|\mathbf{\Delta}\|_F + \frac{2d\|\mathbf{X}\|_2(\|\mathbf{X}\|_2 M + \|\mathbf{X}\|_F)\{2\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F\|\mathbf{X}\mathbf{\Delta}\|_F + \|\mathbf{X}\mathbf{\Delta}\|_F^2\}}{\eta^2} \\
&\le \frac{2d\|\mathbf{X}\|_2^2}{\eta}\|\mathbf{\Delta}\|_F + \frac{2d\|\mathbf{X}\|_2(\|\mathbf{X}\|_2 M + \|\mathbf{X}\|_F)\{2(\|\mathbf{X}\|_2 M + \|\mathbf{X}\|_F)\|\mathbf{X}\|_2\|\mathbf{\Delta}\|_F + \|\mathbf{X}\|_2^2\|\mathbf{\Delta}\|_F^2\}}{\eta^2}.
\end{aligned}
$$

These estimates imply analyticity and finite local Lipschitz constants for all equality and inequality expressions in Appendix A.2 defining $Z^{(1)}$, $Z^{(2)}$, $Y^{(1)}$, and $Y^{(2)}$. For the λ_1-path, fix $\lambda_2 = \lambda_2^0$. On a regime S, s,

$$
\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-T}]_{ij} + 2\lambda_2^0\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji} + \lambda_1 s_{ij} = 0, \qquad (i,j)\in S,
$$

and, when the active Jacobian determinant is nonzero,

$$
\begin{aligned}
&\left[\frac{d\mathbf{W}_{uv}}{d\lambda_1}\right]_{(u,v)\in S} \\
&= -\left(\left[\frac{\partial}{\partial \mathbf{W}_{uv}} \left\{ \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-T}]_{ij} + 2\lambda_2^0\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji} \right\} \right]_{(i,j),(u,v)\in S} \right)^{-1} [s_{ij}]_{(i,j)\in S}.
\end{aligned}
$$

At a support-sign wall for (r,c) with one-sided sign σ,

$$\begin{aligned}
&\mathbf{W}_{rc}=0, \qquad 2\lambda_2^0\mathbf{W}_{rc}(e^{\mathbf{W}\odot\mathbf{W}})_{cr}=0, \qquad \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{rc}+\lambda_1\sigma=0,\\
&\lambda_1=-\sigma\left[\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{rc}\right],\\
&0=\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2^0\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}\\
&\qquad\qquad -\sigma s_{ij}\left[\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{rc}\right], \quad (i,j)\in S,\\
&\left|\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}\right|\le-\sigma\left[\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{rc}\right], \quad (i,j)\notin S\cup\{(r,c)\},
\end{aligned} \tag{39}$$

which is exactly the defining system of $Z^{(1)}_{(r,c),\sigma,S,s}$. If the transition is internal to a regime, then the strict inequalities hold and the active Jacobian determinant vanishes, which is precisely the defining condition of $Y^{(1)}_{S,s}$. Let $\mathcal{T}_{\lambda_1}$ denote the set of transition points in $0<\lambda_1<+\infty$. Consequently,

$$\begin{aligned}
\mathcal{T}_{\lambda_1}\subseteq&\bigcup_{(r,c)\in\{1,\dots,d\}^2}\bigcup_{\sigma=\pm1}\bigcup_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}}\bigcup_{s\in\{\pm1\}^S}Z^{(1)}_{(r,c),\sigma,S,s}\\
&\cup\bigcup_{\varnothing\neq S\subseteq\{1,\dots,d\}^2}\bigcup_{s\in\{\pm1\}^S}Y^{(1)}_{S,s},\\
\#\mathcal{T}_{\lambda_1}\le&\sum_{(r,c)\in\{1,\dots,d\}^2}\sum_{\sigma=\pm1}\sum_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}}\sum_{s\in\{\pm1\}^S}\#Z^{(1)}_{(r,c),\sigma,S,s}+\sum_{\varnothing\neq S\subseteq\{1,\dots,d\}^2}\sum_{s\in\{\pm1\}^S}\#Y^{(1)}_{S,s}.
\end{aligned}$$

Moreover, at $\mathbf{W}=\mathbf{0}$,

$$[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}=\delta_{ij}, \qquad \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}=-\frac{d(\mathbf{X}^\top\mathbf{X})_{ij}}{\|\mathbf{X}\|_F^2}.$$

The zero-branch stationarity condition is equivalent to

$$\mathbf{0}\in\left[\delta_{ij}-\frac{d(\mathbf{X}^\top\mathbf{X})_{ij}}{\|\mathbf{X}\|_F^2}\right]_{i,j=1}^{d}+\lambda_1\partial\|\mathbf{W}\|_1, \qquad \max_{i,j}\left|\delta_{ij}-\frac{d(\mathbf{X}^\top\mathbf{X})_{ij}}{\|\mathbf{X}\|_F^2}\right|\le\lambda_1.$$

Thus, by the definition of $\lambda_{\max}$, the zero branch is stationary for every $\lambda_1\ge\lambda_{\max}$ and contributes no additional support–sign transition beyond $\lambda_{\max}$. For the property in λ_2-path, fix $\lambda_1=\lambda_1^0$. In a regime S,s,

$$\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}+\lambda_1^0 s_{ij}=0, \qquad (i,j)\in S,$$

and, when the active Jacobian determinant is nonzero,

$$\begin{aligned}
\left[\frac{d\mathbf{W}_{uv}}{d\lambda_2}\right]_{(u,v)\in S}=-\Bigg(&\left[\frac{\partial}{\partial\mathbf{W}_{uv}}\left\{\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}+2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}\right\}\right]_{(i,j),(u,v)\in S}\Bigg)^{-1}\\
&\times\left[2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}\right]_{(i,j)\in S}.
\end{aligned}$$

At a support-sign wall for (r,c) with one-sided sign σ,

$$\begin{aligned}
&\mathbf{W}_{rc}=0, \qquad 2\lambda_2\mathbf{W}_{rc}(e^{\mathbf{W}\odot\mathbf{W}})_{cr}=0, \qquad \frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{rc}+\lambda_1^0\sigma=0,\\
&2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji}=-\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}-[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}-\lambda_1^0 s_{ij}, \quad (i,j)\in S,\\
&\left|\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}+[(\mathbf{I}-\mathbf{W})^{-T}]_{ij}\right|\le\lambda_1^0, \quad (i,j)\notin S\cup\{(r,c)\},
\end{aligned} \tag{40}$$

which is exactly the defining system of $Z^{(2)}_{(r,c),\sigma,S,s}$. If the transition is internal to a regime, then the strict inequalities hold and the active Jacobian determinant vanishes, so the transition belongs to $Y^{(2)}_{S,s}$. Let $\mathcal{T}_{\lambda_2}$ denote the set of transition points in $0 < \lambda_2 < \lambda_2^{\max}$. Hence

$$\begin{aligned}\mathcal{T}_{\lambda_2} \subseteq & \bigcup_{(r,c)\in\{1,\dots,d\}^2} \bigcup_{\sigma=\pm 1} \bigcup_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}} \bigcup_{s\in\{\pm1\}^S} Z^{(2)}_{(r,c),\sigma,S,s} \\ & \cup \bigcup_{\varnothing\neq S\subseteq\{1,\dots,d\}^2} \bigcup_{s\in\{\pm1\}^S} Y^{(2)}_{S,s},\end{aligned}$$

$$\#\mathcal{T}_{\lambda_2} \le \sum_{(r,c)\in\{1,\dots,d\}^2} \sum_{\sigma=\pm1} \sum_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}} \sum_{s\in\{\pm1\}^S} \#Z^{(2)}_{(r,c),\sigma,S,s} + \sum_{\varnothing\neq S\subseteq\{1,\dots,d\}^2} \sum_{s\in\{\pm1\}^S} \#Y^{(2)}_{S,s}.$$

Then, it remains to prove that the right-hand sides are finite with probability one. For $Z^{(1)}_{(r,c),\sigma,S,s}$, $Z^{(2)}_{(r,c),\sigma,S,s}$, $Y^{(1)}_{S,s}$, and $Y^{(2)}_{S,s}$, respectively, the numbers of scalar equality constraints are

$$\begin{aligned}&1+1+(d^2-|S|-1)+|S| = d^2+1, \qquad 1+(d^2-|S|-1)+1+|S| = d^2+1,\\ &(d^2-|S|)+|S|+1 = d^2+1, \qquad (d^2-|S|)+|S|+1 = d^2+1.\end{aligned}$$

For each equality system, the Jacobian with respect to the unknown coordinates $(\mathbf{W}_{11}, \mathbf{W}_{12}, \dots, \mathbf{W}_{dd}, \lambda_1)$ or $(\mathbf{W}_{11}, \mathbf{W}_{12}, \dots, \mathbf{W}_{dd}, \lambda_2)$ has analytic entries, with coordinate-zero rows and parameter rows satisfying

$$\frac{\partial \mathbf{W}_{ij}}{\partial \mathbf{W}_{uv}} = \mathbf{1}_{\{i=u,j=v\}}, \qquad \frac{\partial}{\partial\lambda_1}\left\{\lambda_1 + \sigma\left[\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{rc}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-T}]_{rc}\right]\right\} = 1,$$

$$\frac{\partial}{\partial\lambda_2}\left\{\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} + [(\mathbf{I}-\mathbf{W})^{-T}]_{ij} + 2\lambda_2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji} + \lambda_1^0 s_{ij}\right\} = 2\mathbf{W}_{ij}(e^{\mathbf{W}\odot\mathbf{W}})_{ji},$$

and the data derivatives entering the residual-gradient rows are

$$\begin{aligned}\frac{\partial}{\partial X_{\alpha\beta}}\left[\frac{d[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2}\right] &= d\frac{\mathbf{1}_{\{i=\beta\}}(\mathbf{X}\mathbf{W}-\mathbf{X})_{\alpha j} + X_{\alpha i}(\mathbf{W}_{\beta j}-\mathbf{1}_{\{\beta=j\}})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2} \\ &\quad - d\frac{[\mathbf{X}^\top(\mathbf{X}\mathbf{W}-\mathbf{X})]_{ij}\cdot 2\sum_{m=1}^d(\mathbf{X}-\mathbf{X}\mathbf{W})_{\alpha m}(\mathbf{1}_{\{\beta=m\}}-\mathbf{W}_{\beta m})}{\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^4}.\end{aligned}$$

For each fixed configuration, rank failure at a compact-domain solution is contained in the event that all equality left-hand sides vanish and all maximal unknown-Jacobian minors vanish; these are analytic equations in $(\mathbf{W}, \lambda_1, \mathbf{X})$ or $(\mathbf{W}, \lambda_2, \mathbf{X})$. Analytic parametric transversality, together with absolute continuity of $\mathrm{vec}(\mathbf{X})$, gives

$$\begin{aligned}&\mathbb{P}\left(\exists(\mathbf{W},\lambda_1)\in Z^{(1)}_{(r,c),\sigma,S,s} : \operatorname{rank}\frac{\partial(\text{all equality left-hand sides})}{\partial(\mathbf{W},\lambda_1)} < d^2+1\right) = 0,\\ &\mathbb{P}\left(\exists(\mathbf{W},\lambda_2)\in Z^{(2)}_{(r,c),\sigma,S,s} : \operatorname{rank}\frac{\partial(\text{all equality left-hand sides})}{\partial(\mathbf{W},\lambda_2)} < d^2+1\right) = 0,\\ &\mathbb{P}\left(\exists(\mathbf{W},\lambda_1)\in Y^{(1)}_{S,s} : \operatorname{rank}\frac{\partial(\text{all equality left-hand sides})}{\partial(\mathbf{W},\lambda_1)} < d^2+1\right) = 0,\\ &\mathbb{P}\left(\exists(\mathbf{W},\lambda_2)\in Y^{(2)}_{S,s} : \operatorname{rank}\frac{\partial(\text{all equality left-hand sides})}{\partial(\mathbf{W},\lambda_2)} < d^2+1\right) = 0.\end{aligned} \tag{41}$$

Since

$$\sum_{(r,c)\in\{1,\dots,d\}^2}\sum_{\sigma=\pm1}\sum_{S\subseteq\{1,\dots,d\}^2\setminus\{(r,c)\}}\sum_{s\in\{\pm1\}^S} 1 = 2d^2\sum_{q=0}^{d^2-1}\binom{d^2-1}{q}2^q = 2d^2\,3^{d^2-1},$$

and

$$\sum_{\varnothing\neq S\subseteq\{1,\dots,d\}^2}\sum_{s\in\{\pm1\}^S} 1 = \sum_{q=1}^{d^2}\binom{d^2}{q}2^q = 3^{d^2}-1,$$

the probability that any of the finitely many $Z^{(1)}, Z^{(2)}, Y^{(1)}, Y^{(2)}$ systems loses full-rank isolatedness at a compact-domain solution is $\left\{2 \cdot 2d^2\, 3^{d^2-1} + 2(3^{d^2}-1)\right\} \cdot 0 = 0$. On the complementary event, every compact-domain solution of every square system (39)-(40) is isolated. If one of these sets were infinite, compactness would give distinct solutions $(\mathbf{W}_m, \lambda_m) \to (\mathbf{W}_*, \lambda_*)$ via (38), continuity would give that $(\mathbf{W}_*, \lambda_*)$ solves the same equality system. Let $\mathcal{E}_{\mathrm{eq}}$ denote the vector of all equality left-hand sides in this fixed square system. Full rank of the equality Jacobian at the limit (41) would imply, by the inverse function theorem, that for some $\rho > 0$,

$$\{(\mathbf{W}, \lambda) : \mathcal{E}_{\mathrm{eq}}(\mathbf{W}, \lambda) = \mathbf{0}, \quad \|(\mathbf{W}, \lambda) - (\mathbf{W}_*, \lambda_*)\|_2 < \rho\} = \{(\mathbf{W}_*, \lambda_*)\},$$

contradicting the convergence of distinct solutions. Therefore, with probability one,

$$\#Z^{(1)}_{(r,c),\sigma,S,s} < \infty, \qquad \#Z^{(2)}_{(r,c),\sigma,S,s} < \infty, \qquad \#Y^{(1)}_{S,s} < \infty, \qquad \#Y^{(2)}_{S,s} < \infty$$

for every admissible configuration. Away from those finitely many points, the support-sign inequalities are strict and the active Jacobian determinant is nonzero, so the implicit function theorem yields real-analytic one-dimensional charts. Since

$$\sum_{S \subseteq \{1,\dots,d\}^2} \sum_{s \in \{\pm 1\}^S} 1 = \sum_{q=0}^{d^2} \binom{d^2}{q} 2^q = 3^{d^2},$$

there are finitely many support-sign regimes. Hence the exact stationary solution graph is a finite union of analytic arcs and finitely many Z-wall and Y-singular transition points, which is a finite analytic stratification, and the two asserted cardinality bounds follow. □

M Experimental Details

M.1 Dataset Details

Synthetic Data. We generate synthetic data from linear SEMs. For each run, we first generate a master weighted DAG $\mathbf{W}^\star \in \mathbb{R}^{D \times D}$ and a master data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. Experiments with different sample sizes and graph sizes are obtained by taking subsets of rows and variables from this master matrix. Unless otherwise stated, variables are taken according to the sampled topological order. Given $\mathbf{W}^\star$, data are generated by $\mathbf{X} = \mathbf{N}(\mathbf{I} - \mathbf{W}^\star)^{-1}$, where the columns of $\mathbf{N}$ are mutually independent. After generation, each column of $\mathbf{X}$ is centered and scaled to empirical variance one.

For all DGPs, a topological order π is sampled uniformly at random from all permutations of $[D]$. Edges are only allowed from earlier to later nodes in this order. After all edge weights are assigned, we apply the same column-wise stabilization step: for every column j, if $\sum_i |\mathbf{W}^\star_{ij}| > 0.9$, then all incoming weights of node j are multiplied by $0.9/\sum_i |\mathbf{W}^\star_{ij}|$. This preserves the DAG support and prevents unstable SEMs.

DGP-1: degree-controlled ER SEM. Let s_{er} be the target expected degree. For every ordered pair $\pi(a) \to \pi(b)$ with $a < b$, include the edge independently with probability $p_{\mathrm{er}} = s_{\mathrm{er}}/(D-1)$. Conditional on inclusion, sample

$$\mathbf{W}^\star_{\pi(a),\pi(b)} = S_{\pi(a),\pi(b)} U_{\pi(a),\pi(b)}, \qquad S_{\pi(a),\pi(b)} \sim \mathrm{Unif}(\{-1,1\}), \quad U_{\pi(a),\pi(b)} \sim \mathrm{Unif}(0.5, 2.0).$$

The noise variables are independent Gaussian, $\mathbf{N}_{mj} \sim \mathcal{N}(0,1)$. We use $s_{\mathrm{er}} \in \{2,4\}$ in the main experiments.

DGP-2: modular scale-free heteroskedastic SEM. Partition the nodes into $B = 5$ consecutive modules under the topological order, with module sizes differing by at most one. For each node $\pi(b)$, let $P_b = \{\pi(a) : a < b\}$. Its number of parents is $M_b = \min\{b-1, 1 + \mathrm{Pois}(s_{\mathrm{sf}})\}$. Conditional on M_b, parents are sampled without replacement from P_b with probabilities proportional to

$$q_{a,b} = \big(1+\mathrm{outdeg}(\pi(a))\big)^{1.5}\big(\mathbb{I}\{\pi(a), \pi(b) \text{ in the same module}\} + 0.25\,\mathbb{I}\{\pi(a), \pi(b) \text{ in different modules}\}\big).$$

For each selected parent $\pi(a)$, sample

$$\mathbf{W}^\star_{\pi(a),\pi(b)} = S_{\pi(a),\pi(b)}\big[B_{\pi(a),\pi(b)} U^{(1)}_{\pi(a),\pi(b)} + (1 - B_{\pi(a),\pi(b)}) U^{(2)}_{\pi(a),\pi(b)}\big],$$

where $S_{\pi(a),\pi(b)} \sim \mathrm{Unif}(\{-1,1\})$, $B_{\pi(a),\pi(b)} \sim \mathrm{Bernoulli}(0.7)$, $U^{(1)}_{\pi(a),\pi(b)} \sim \mathrm{Unif}(0.8,2.0)$, and $U^{(2)}_{\pi(a),\pi(b)} \sim \mathrm{Unif}(0.2,0.5)$. The noise is heteroskedastic Laplace:

$$\mathbf{N}_{mj} = \tau_j Z_{mj}, \qquad \log \tau_j \sim \mathrm{Unif}(-\log 2, \log 2), \qquad Z_{mj} \sim \mathrm{Laplace}(0, 1/\sqrt{2}).$$

Thus Z_{mj} has unit variance. We use $s_{\mathrm{sf}} \in \{1,2\}$.

DGP-3: near-collinear transition-rich SEM. This DGP is generated in three steps. First, for every pair $(\pi(2r-1), \pi(2r))$, include the copy edge $\pi(2r-1) \to \pi(2r)$ with weight 0.8, i.e.,

$$\mathbf{W}^{\star}_{\pi(2r-1),\pi(2r)} = 0.8, \qquad r = 1, \ldots, \lfloor D/2 \rfloor.$$

Second, for every pair index r and every node $\pi(b)$ with $b > 2r$, draw $C_{r,b} \sim \mathrm{Bernoulli}(p_{\mathrm{pair}})$, where $p_{\mathrm{pair}} = \min\{4/D, 0.2\}$. If $C_{r,b} = 1$, include both edges $\pi(2r-1) \to \pi(b)$ and $\pi(2r) \to \pi(b)$, with

$$\mathbf{W}^{\star}_{\pi(2r-1),\pi(b)} = S_{r,b} U_{r,b}, \qquad \mathbf{W}^{\star}_{\pi(2r),\pi(b)} = -S_{r,b}(U_{r,b} + \Delta_{r,b}),$$

where $S_{r,b} \sim \mathrm{Unif}(\{-1,1\})$, $U_{r,b} \sim \mathrm{Unif}(0.4,1.0)$, and $\Delta_{r,b} \sim \mathrm{Unif}(-0.1,0.1)$. Third, for every remaining ordered pair $\pi(a) \to \pi(b)$ with $a < b$ not already selected, include a weak edge independently with probability $p_{\mathrm{weak}} = \min\{2/D, 0.1\}$ and weight

$$\mathbf{W}^{\star}_{\pi(a),\pi(b)} = S_{\pi(a),\pi(b)} U_{\pi(a),\pi(b)}, \qquad S_{\pi(a),\pi(b)} \sim \mathrm{Unif}(\{-1,1\}), \quad U_{\pi(a),\pi(b)} \sim \mathrm{Unif}(0.15,0.35).$$

The noises are independent Gaussian with

$$\mathbf{N}_{mj} \sim \mathcal{N}(0, \tau_j^2), \qquad \tau_{\pi(2r-1)} = 1, \quad \tau_{\pi(2r)} = 0.25, \quad \tau_j = 1 \text{ otherwise.}$$

All random variables appearing above are mutually independent unless explicitly coupled by the construction.

Real and Public Benchmarks. Sachs contains single-cell flow-cytometry measurements of protein-signaling molecules. We use the continuous curated release and take the standard `cd3cd28.csv` condition, which gives $n = 853$ samples and $d = 11$ variables. We retain only the 11 measured signaling variables, remove non-numeric metadata if present, and use the released signaling network only as a reference graph. DREAM4-10 consists of five 10-gene in-silico networks from the DREAM4 benchmark. We use the Bioconductor `DREAM4` release, where each network is stored as a 10×136 assay matrix; we transpose it to match our row-wise convention, giving $n = 136$ and $d = 10$. DREAM4-100MF is the 100-gene multifactorial DREAM4 benchmark. We use only the multifactorial steady-state matrices, giving $n = 100$ perturbation conditions and $d = 100$ genes, and do not use the time-series, knockout, or knockdown files. SynTReN-300 is the fixed SynTReN-generated dataset distributed through `grndata`; `syntren300.data` contains $n = 800$ simulated microarray samples over $d = 300$ genes, and `syntren300.net` provides the corresponding reference network. These benchmarks are used to evaluate computational efficiency and stationarity control of the regularization path, rather than to claim that a linear SEM is the exact data-generating model.

M.2 Implementation

Batch Solvers At a grid value λ_m, the batch solver was initialized from the raw matrix returned at λ_{m-1}, without thresholding. Unless otherwise stated, all stationarity gaps reported in the paper were recomputed after optimization by a separate KKT-residual evaluator, rather than taken from the stopping criterion of the optimizer.

For NOTEARS [Zheng et al., 2018], the Grid-WS batch solver is a MATLAB implementation of the standard augmented-Lagrangian method with a bound-constrained limited-memory BFGS inner solver. The ℓ_1 penalty is handled by the splitting $\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$, $\mathbf{W}^+, \mathbf{W}^- \geq 0$, with diagonal entries fixed at zero. The augmented-Lagrangian parameters are initialized as $\rho_0 = 1$ and $\alpha_0 = 0$, with update $\alpha \leftarrow \alpha + \rho h(\mathbf{W})$ and $\rho \leftarrow 10\rho$ whenever the acyclicity residual fails to decrease by at least a factor of 0.25. We use $\rho_{\max} = 10^{16}$, at most 100 outer iterations, and at most 2000 L-BFGS iterations per inner problem. The inner solver uses analytic gradients and is terminated when the projected-gradient norm is below 10^{-9} or the relative objective decrease is below 10^{-12}. A grid point is declared converged when $|h(\mathbf{W})| \leq 10^{-8}$ and the normalized KKT gap is below 10^{-6}.

For GOLEM [Ng et al., 2020], the Grid-WS batch solver first performs a warm-started Adam stage and then applies an active-set quasi-Newton polishing step. Adam is run with learning rate 10^{-3}, momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, and at most 5×10^4 iterations; it is stopped early if the relative objective decrease over 500 consecutive iterations is below 10^{-8}. The active set for polishing is initialized as $\mathcal{A} = \{(i,j) : |\mathbf{W}_{ij}| > 10^{-6}\}$. On this active set, signs are fixed and a limited-memory BFGS solver is run for at most 1000 iterations, with gradient tolerance 10^{-8}. After polishing, inactive KKT violations are checked; coordinates with violation larger than 10^{-6} are added to the active set, and the polishing step is repeated for at most 5 active-set rounds. Trial points with $\sigma_{\min}(\mathbf{I} - \mathbf{W}) < 10^{-8}$ are rejected in the line search.

For DAGMA [Bello et al., 2022], we use the same augmented-Lagrangian/L-BFGS framework as for NOTEARS, but replace the matrix-exponential acyclicity function by the log-determinant certificate $h^s_{\mathrm{ldet}}(\mathbf{W})$ with $s = 1$. The ℓ_1 penalty is again handled by the splitting $\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$, and the previous grid solution is used as the initialization. The augmented-Lagrangian parameters are $\rho_0 = 1$, $\alpha_0 = 0$, $\rho_{\max} = 10^{16}$, with the same $\rho \leftarrow 10\rho$ update rule as in NOTEARS. Each inner problem is solved by bound-constrained L-BFGS for at most 2000 iterations, using analytic gradients. The log-determinant domain condition $s\mathbf{I} - \mathbf{W} \circ \mathbf{W} \succ 0$ is enforced by Cholesky checking; a trial point is rejected if the Cholesky factorization fails or the smallest diagonal pivot is below 10^{-10}. The solver stops when $|h^s_{\mathrm{ldet}}(\mathbf{W})| \le 10^{-8}$ and the normalized KKT gap is below 10^{-6}, or when the maximum number of outer iterations is reached.

Comparisons All experiments are implemented in MATLAB R2021a and executed on a workstation running Windows 10 with 32GB RAM and an Intel Core i7 CPU. In all subfigures of Figure 2, Figure 3, and Figure 4, we fix $d = 160$ for the time-vs. n plots and $n = 5000$ for the time-vs. d plots. In Table 1, Table 3, and Table 4, the random seed is used only to fix the generated synthetic datasets.

Test NLL Metric We use three data splits: $\mathbf{X}^{\mathrm{tr}}$, $\mathbf{X}^{\mathrm{val}}$, and $\mathbf{X}^{\mathrm{te}}$. The training split is used to compute the regularization path and to refit candidate graphs, the validation split is used only to select one graph from the candidate set, and the test split is used only once for final reporting. Let $\mathcal{C}$ denote the set of candidate DAG supports produced by either `DAG-flow` or a warm-start grid. For each $G \in \mathcal{C}$, we refit an unregularized linear SEM on the training split restricted to the support G:

$$\widehat{\mathbf{W}}_G \in \arg\min_{\mathrm{supp}(\mathbf{W}) \subseteq G} \frac{1}{2n_{\mathrm{tr}}} \left\| \mathbf{X}^{\mathrm{tr}} - \mathbf{X}^{\mathrm{tr}}\mathbf{W} \right\|_F^2 .$$

This refitting step removes shrinkage bias from the ℓ_1-penalized weights and makes all candidate graphs comparable under the same SEM score. For each node j, we estimate the residual variance by

$$\widehat{\sigma}_j^2(G) = \frac{1}{n_{\mathrm{tr}}} \left\| \mathbf{X}^{\mathrm{tr}}_{:,j} - \mathbf{X}^{\mathrm{tr}}\widehat{\mathbf{W}}_{G,:,j} \right\|_2^2 .$$

For a held-out split $D \in \{\mathrm{val}, \mathrm{te}\}$, define the average Gaussian SEM negative log-likelihood as

$$\mathrm{NLL}_D(G) = \frac{1}{2d} \sum_{j=1}^{d} \left[\log\left(2\pi\widehat{\sigma}_j^2(G)\right) + \frac{\left\| \mathbf{X}^D_{:,j} - \mathbf{X}^D\widehat{\mathbf{W}}_{G,:,j} \right\|_2^2}{n_D \widehat{\sigma}_j^2(G)} \right] .$$

The selected graph is $\widehat{G}_{\mathcal{C}} \in \arg\min_{G \in \mathcal{C}} \mathrm{NLL}_{\mathrm{val}}(G)$, and the reported score is

$$\mathrm{TestNLL}(\mathcal{C}) = \mathrm{NLL}_{\mathrm{te}}(\widehat{G}_{\mathcal{C}}).$$

If the least-squares refit is rank-deficient, we use the minimum-norm least-squares solution.

N Additional Empirical Results

In this section, we present additional experimental results.

Path Exactness and Scalability. Figure 3 and Table 3 report the accuracy and runtime on DGP2, while Figure 4 and Table 4 report the corresponding results on DGP3. The experimental setup and our observations are consistent with those presented in the main body.

We further conduct an ablation study on different grid partitioning schemes, with results presented in Figure 5. Our method continues to demonstrate a significant advantage across all settings.

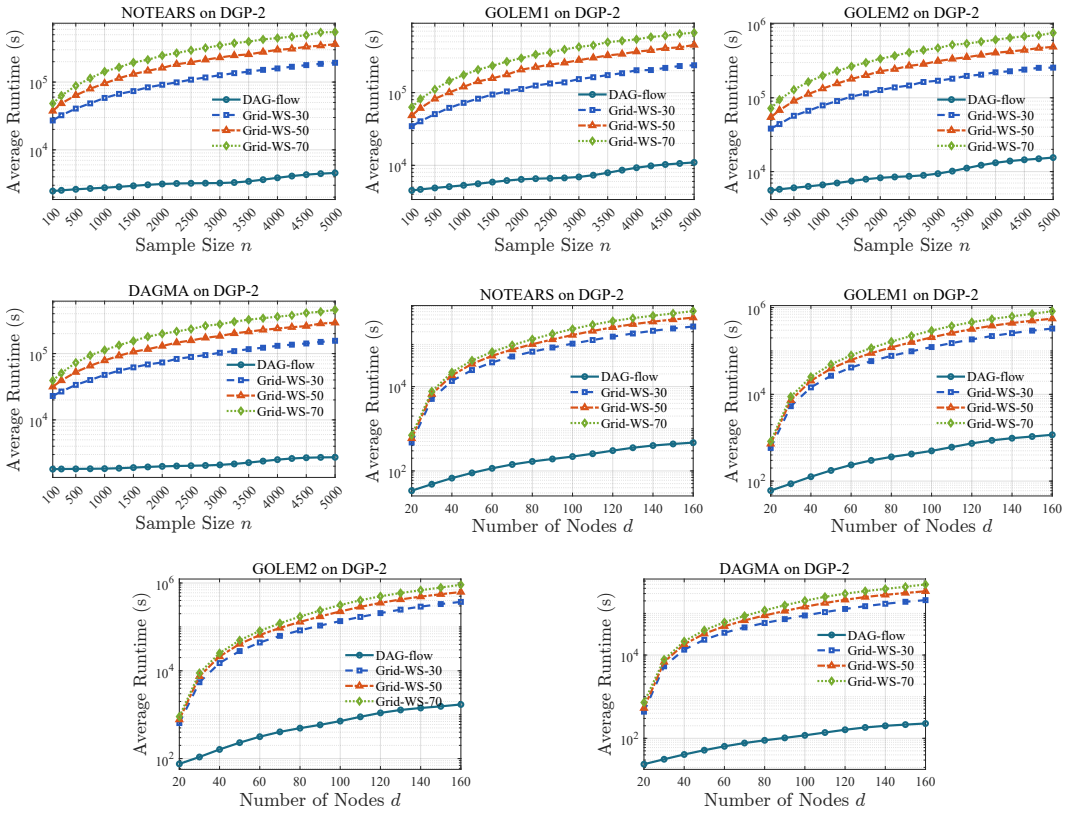


Figure 3. Average wall-clock time for computing the entire solution path, averaged over 20 runs. Subplot titles indicate the experimental settings; when one of n and d is varied, the other dimension is held fixed.

Edge-Level Stability As just one illustrative example, we use edge persistence to summarize the stability information contained in a full regularization path. For an edge e, its persistence π_e is the fraction of the evaluated regularization range over which e is active. For `DAG-flow`, π_e is computed from the continuous path segments; for Grid-WS-G, it is computed as the fraction of the G grid solutions in which e appears. Edges are ranked by π_e, and T_K denotes the top-K ranked edges.

We report three stability diagnostics. $\mathrm{Persist@}K$ is the average persistence of the edges in T_K, so larger values mean that the selected edges survive over broader tuning regimes. $\mathrm{Stab@}K$ is the average pairwise Jaccard similarity of T_K across random train, validation, test splits, measuring whether the same high-persistence edges are recovered reproducibly. $\mathrm{Fragile@}K$ is the fraction of edges in T_K with $\pi_e < 0.12$; smaller values mean that the ranking is not dominated by edges that appear only in very narrow tuning intervals. These metrics do not require a ground-truth graph and are used only to evaluate stability of the learned candidate edge set.

The results in Table 5 show that continuous path information improves edge rankings beyond what is available from a finite grid. On Sachs, the gain is moderate because the problem is small and many stable edges are already visible to Grid-WS-G. On DREAM4-10, the grid solutions are more sensitive to the split, leading to a larger gap in $\mathrm{Stab@}K$. For DREAM4-100MF, both methods are less stable because $n \approx d$, but `DAG-flow` still reduces the fraction of fragile top-ranked edges. On SynTReN-300, the larger sample size makes persistence estimation more reliable, and `DAG-flow` achieves the strongest improvement in $\mathrm{Persist@}K$. Overall, the path representation yields edge rankings whose leading edges persist over broader tuning ranges and are more reproducible across data splits.

Table 3. Active stationarity residuals of `DAG-flow` and Grid-WS-1 on DGP-2 at the finite-time cutoff (lower is better). For each row, the initial solution is obtained at the listed λ, and the target parameter is $\lambda + 0.10\,\lambda_{\max}$. The rightmost column reports the average percentage of transition points that triggered a warm-restart in path-following. We set $\lambda_2 = 0.02$ for GOLEM1 and $\lambda_1 = 0.005$ for GOLEM2. Values below 10^{-6} are reported as 0.

Initial λ	Seed	NOTEARS		GOLEM1		GOLEM2		DAGMA		Avg. Restart
		`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	
$0.10\lambda_{\max}$		0.00	1.18(0.213)	0.00	0.519(0.069)	0.00	0.617(0.092)	0.00	1.09(0.183)	21%
$0.15\lambda_{\max}$		0.00	1.04(0.154)	0.00	0.478(0.041)	0.00	0.571(0.067)	0.00	0.998(0.136)	25%
$0.20\lambda_{\max}$		0.00	0.923(0.116)	0.00	0.452(0.083)	0.00	0.548(0.119)	0.00	0.887(0.151)	23%
$0.25\lambda_{\max}$		0.00	0.861(0.173)	0.00	0.397(0.046)	0.00	0.482(0.058)	0.00	0.832(0.097)	30%
$0.30\lambda_{\max}$		0.00	0.792(0.099)	0.00	0.424(0.094)	0.00	0.463(0.074)	0.00	0.758(0.128)	27%
$0.35\lambda_{\max}$		0.00	0.846(0.210)	0.00	0.356(0.035)	0.00	0.511(0.136)	0.00	0.804(0.196)	32%
$0.40\lambda_{\max}$	42	0.00	0.674(0.081)	0.00	0.331(0.052)	0.00	0.418(0.046)	0.00	0.632(0.072)	28%
$0.45\lambda_{\max}$		0.00	0.602(0.118)	0.00	0.319(0.037)	0.00	0.391(0.075)	0.00	0.584(0.102)	29%
$0.50\lambda_{\max}$		0.00	0.523(0.076)	0.00	0.274(0.031)	0.00	0.342(0.038)	0.00	0.506(0.064)	24%
$0.55\lambda_{\max}$		0.00	0.479(0.104)	0.00	0.251(0.046)	0.00	0.318(0.063)	0.00	0.451(0.088)	22%
$0.60\lambda_{\max}$		0.00	0.413(0.062)	0.00	0.232(0.026)	0.00	0.287(0.032)	0.00	0.421(0.057)	20%
$0.65\lambda_{\max}$		0.00	0.366(0.048)	0.00	0.218(0.031)	0.00	0.264(0.047)	0.00	0.351(0.044)	18%
$0.70\lambda_{\max}$		0.00	0.327(0.055)	0.00	0.198(0.023)	0.00	0.241(0.029)	0.00	0.318(0.060)	17%
$0.10\lambda_{\max}$		0.00	1.25(0.168)	0.00	0.487(0.048)	0.00	0.648(0.081)	0.00	1.16(0.211)	19%
$0.15\lambda_{\max}$		0.00	0.981(0.203)	0.00	0.511(0.089)	0.00	0.529(0.123)	0.00	0.946(0.126)	23%
$0.20\lambda_{\max}$		0.00	0.954(0.112)	0.00	0.406(0.037)	0.00	0.572(0.066)	0.00	0.902(0.176)	26%
$0.25\lambda_{\max}$		0.00	0.904(0.142)	0.00	0.432(0.077)	0.00	0.504(0.095)	0.00	0.861(0.112)	28%
$0.30\lambda_{\max}$		0.00	0.731(0.086)	0.00	0.389(0.045)	0.00	0.437(0.049)	0.00	0.705(0.092)	31%
$0.35\lambda_{\max}$		0.00	0.873(0.232)	0.00	0.341(0.061)	0.00	0.472(0.108)	0.00	0.825(0.220)	27%
$0.40\lambda_{\max}$	2026	0.00	0.706(0.118)	0.00	0.358(0.092)	0.00	0.447(0.069)	0.00	0.668(0.080)	29%
$0.45\lambda_{\max}$		0.00	0.557(0.063)	0.00	0.296(0.030)	0.00	0.368(0.041)	0.00	0.531(0.089)	25%
$0.50\lambda_{\max}$		0.00	0.548(0.123)	0.00	0.302(0.055)	0.00	0.359(0.082)	0.00	0.519(0.117)	24%
$0.55\lambda_{\max}$		0.00	0.434(0.055)	0.00	0.238(0.022)	0.00	0.301(0.033)	0.00	0.418(0.050)	21%
$0.60\lambda_{\max}$		0.00	0.438(0.083)	0.00	0.249(0.040)	0.00	0.294(0.052)	0.00	0.407(0.071)	18%
$0.65\lambda_{\max}$		0.00	0.342(0.042)	0.00	0.207(0.026)	0.00	0.255(0.031)	0.00	0.337(0.039)	18%
$0.70\lambda_{\max}$		0.00	0.349(0.074)	0.00	0.218(0.033)	0.00	0.249(0.045)	0.00	0.333(0.065)	16%

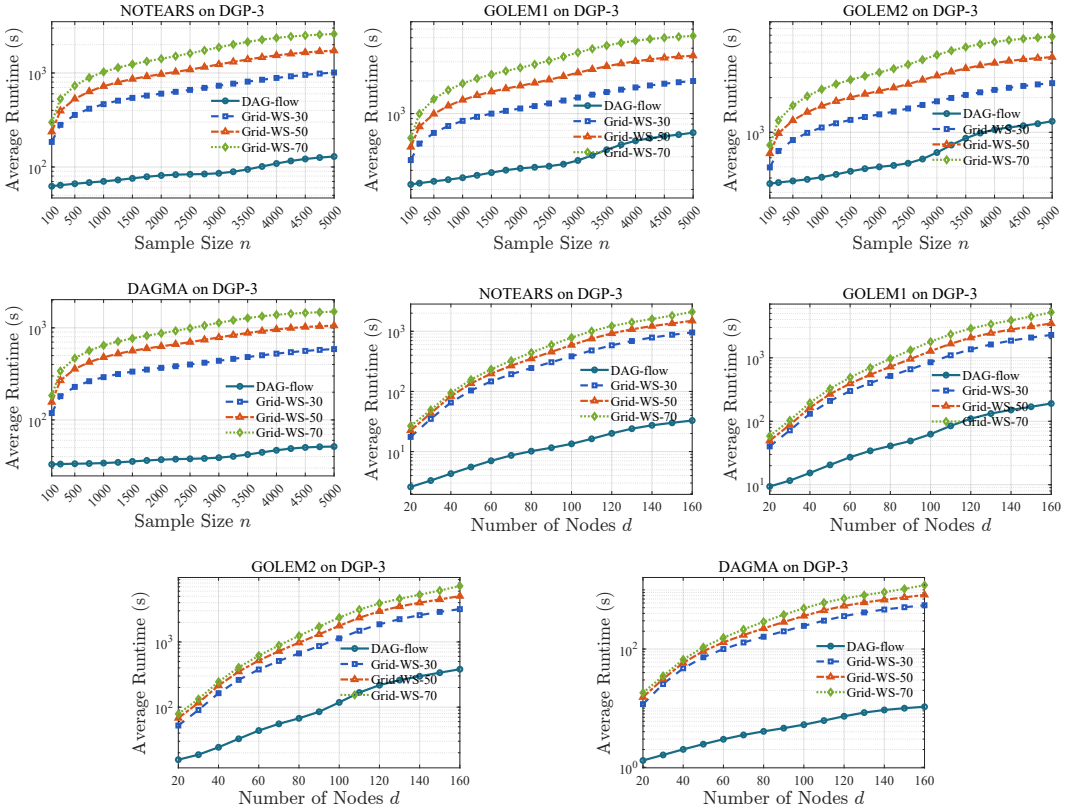


Figure 4. Average wall-clock time for computing the entire solution path, averaged over 20 runs. Subplot titles indicate the experimental settings; when one of n and d is varied, the other dimension is held fixed.

Table 4. Active stationarity residuals of `DAG-flow` and Grid-WS-1 on DGP-3 at the finite-time cutoff (lower is better). For each row, the initial solution is obtained at the listed λ, and the target parameter is $\lambda + 0.10\,\lambda_{\max}$. The rightmost column reports the average percentage of transition points that triggered a warm-restart in path-following. We set $\lambda_2 = 0.02$ for GOLEM1 and $\lambda_1 = 0.005$ for GOLEM2. Values below 10^{-6} are reported as 0.

Initial λ	Seed	NOTEARS		GOLEM1		GOLEM2		DAGMA		Avg. Restart
		`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	`DAG-flow`	Grid-WS	
$0.10\lambda_{\max}$	42	0.00	1.62(0.311)	0.00	0.712(0.102)	0.00	0.835(0.168)	0.00	1.51(0.260)	27%
$0.15\lambda_{\max}$		0.00	1.49(0.184)	0.00	0.654(0.089)	0.00	0.802(0.107)	0.00	1.43(0.215)	30%
$0.20\lambda_{\max}$		0.00	1.35(0.267)	0.00	0.691(0.141)	0.00	0.744(0.096)	0.00	1.29(0.173)	29%
$0.25\lambda_{\max}$		0.00	1.41(0.352)	0.00	0.588(0.074)	0.00	0.778(0.180)	0.00	1.34(0.286)	33%
$0.30\lambda_{\max}$		0.00	1.22(0.156)	0.00	0.603(0.117)	0.00	0.696(0.083)	0.00	1.18(0.151)	31%
$0.35\lambda_{\max}$		0.00	1.31(0.302)	0.00	0.541(0.068)	0.00	0.713(0.153)	0.00	1.24(0.255)	34%
$0.40\lambda_{\max}$		0.00	0.986(0.128)	0.00	0.505(0.087)	0.00	0.591(0.069)	0.00	0.942(0.119)	28%
$0.45\lambda_{\max}$		0.00	0.874(0.181)	0.00	0.472(0.055)	0.00	0.557(0.105)	0.00	0.851(0.162)	26%
$0.50\lambda_{\max}$		0.00	0.751(0.096)	0.00	0.414(0.077)	0.00	0.486(0.061)	0.00	0.734(0.090)	24%
$0.55\lambda_{\max}$		0.00	0.668(0.143)	0.00	0.386(0.049)	0.00	0.452(0.092)	0.00	0.649(0.121)	21%
$0.60\lambda_{\max}$		0.00	0.574(0.078)	0.00	0.341(0.063)	0.00	0.418(0.050)	0.00	0.558(0.071)	22%
$0.65\lambda_{\max}$		0.00	0.509(0.102)	0.00	0.318(0.041)	0.00	0.369(0.074)	0.00	0.493(0.097)	18%
$0.70\lambda_{\max}$		0.00	0.467(0.065)	0.00	0.289(0.036)	0.00	0.342(0.045)	0.00	0.456(0.080)	17%
$0.10\lambda_{\max}$	2026	0.00	1.55(0.241)	0.00	0.758(0.126)	0.00	0.811(0.092)	0.00	1.59(0.318)	26%
$0.15\lambda_{\max}$		0.00	1.64(0.334)	0.00	0.621(0.071)	0.00	0.866(0.184)	0.00	1.47(0.205)	28%
$0.20\lambda_{\max}$		0.00	1.28(0.149)	0.00	0.732(0.152)	0.00	0.721(0.082)	0.00	1.36(0.234)	32%
$0.25\lambda_{\max}$		0.00	1.46(0.218)	0.00	0.572(0.060)	0.00	0.794(0.145)	0.00	1.31(0.164)	30%
$0.30\lambda_{\max}$		0.00	1.17(0.204)	0.00	0.628(0.108)	0.00	0.662(0.071)	0.00	1.20(0.199)	34%
$0.35\lambda_{\max}$		0.00	1.36(0.279)	0.00	0.526(0.082)	0.00	0.738(0.160)	0.00	1.16(0.132)	31%
$0.40\lambda_{\max}$		0.00	1.04(0.167)	0.00	0.491(0.056)	0.00	0.612(0.113)	0.00	0.973(0.147)	29%
$0.45\lambda_{\max}$		0.00	0.842(0.097)	0.00	0.508(0.096)	0.00	0.528(0.064)	0.00	0.883(0.182)	27%
$0.50\lambda_{\max}$		0.00	0.792(0.163)	0.00	0.398(0.044)	0.00	0.515(0.104)	0.00	0.701(0.086)	23%
$0.55\lambda_{\max}$		0.00	0.631(0.081)	0.00	0.404(0.079)	0.00	0.431(0.052)	0.00	0.668(0.137)	24%
$0.60\lambda_{\max}$		0.00	0.603(0.130)	0.00	0.329(0.038)	0.00	0.443(0.091)	0.00	0.529(0.063)	20%
$0.65\lambda_{\max}$		0.00	0.487(0.057)	0.00	0.336(0.066)	0.00	0.361(0.041)	0.00	0.513(0.109)	19%
$0.70\lambda_{\max}$		0.00	0.492(0.114)	0.00	0.276(0.034)	0.00	0.359(0.071)	0.00	0.435(0.052)	18%

Table 5. Edge-level stability of persistence-based edge rankings. Grid-WS-G uses $G = 50$. Results are averaged over NOTEARS, GOLEM (λ_1-path), and DAGMA across 20 random splits. Values are shown as mean(std).

Dataset	Method	Top-K	Persist@K ↑	Stab@K ↑	Fragile@K ↓
Sachs	`DAG-flow`	20	$0.643(0.054)$	$0.604(0.079)$	$0.060(0.043)$
	Grid-WS-G	20	$0.552(0.071)$	$0.493(0.094)$	$0.142(0.069)$
DREAM4-10	`DAG-flow`	20	$0.511(0.086)$	$0.423(0.107)$	$0.176(0.084)$
	Grid-WS-G	20	$0.386(0.096)$	$0.276(0.101)$	$0.358(0.121)$
DREAM4-100MF	`DAG-flow`	50	$0.337(0.052)$	$0.234(0.061)$	$0.316(0.087)$
	Grid-WS-G	50	$0.294(0.049)$	$0.184(0.054)$	$0.443(0.099)$
SynTReN-300	`DAG-flow`	100	$0.462(0.041)$	$0.391(0.052)$	$0.153(0.064)$
	Grid-WS-G	100	$0.344(0.047)$	$0.278(0.058)$	$0.322(0.089)$

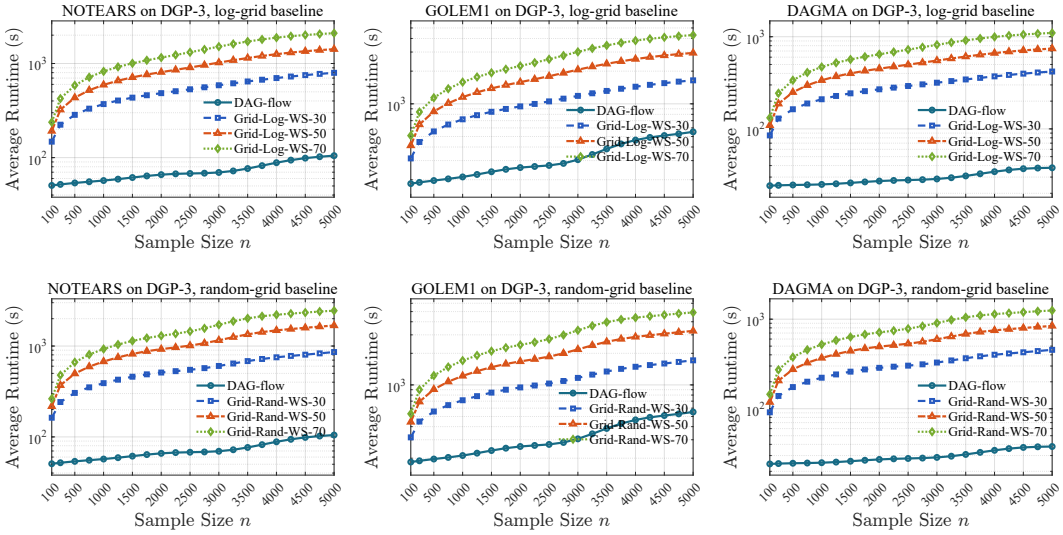


Figure 5. Grid-selection ablation experiments. Average runtime is plotted against sample size n for NOTEARS, GOLEM1, and DAGMA under DGP-3, with $d = 160$ fixed. Each curve is averaged over 20 repetitions. The top three panels use *logarithmically* spaced warm-start grids and the bottom three panels use *randomly* selected warm-start grids, with $G \in \{30, 50, 70\}$ grid points. The results show that the efficiency gain of `DAG-flow` is largely insensitive to whether the grid baseline uses log-spaced or random grid points. This suggests that the computational advantage arises from continuous path following rather than from a favorable choice of discrete grid locations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims we made are accurate.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are in Section 2. The verified proofs appear in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The contribution is primarily a new algorithm. See Appendix B and Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See our supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 7 and Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 7 and Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational research and not tied to particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 7 and Appendix M. Licenses are available referring to the provided citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.