Learning a Forward Operator from Cross-Modal Second-Order Statistics

Anonymous authors
Paper under double-blind review

Abstract

We consider the linear forward model $y = Ax + \epsilon$, $A \in \mathbb{R}^{C\times N}$, $C < N$, with the high-dimensional source $x \in \mathbb{R}^N$ unobserved during training; we learn A from matching the y-covariance $\Sigma_y \in \mathbb{R}^{C\times C}$ together with a given covariance $\Sigma^{\text{ref}} \in \mathbb{R}^{P\times P}$ for Πx, where Π aggregates x into $P < N$ dimensions. The two covariances of different dimensions are linked through the pseudo-inverse of A: With $B := \Pi A^\dagger$, A can be learned under the covariance-matching objective $B\Sigma_y B^T \approx \Sigma^{\text{ref}}$. We parameterise A in low-rank (LoRA) form: $A = A_0 + UV^T$, with A_0 a known baseline. Our theoretical analysis addresses three questions relevant to training under this objective: identifiability of the matching set in B-space, the size of the derivative of $A \mapsto A^\dagger$ through which the gradient passes, and conditions under which training keeps A close to a warm-start A_0. Identifiability tells us how the set of B that best fits the references sits in $\mathbb{R}^{P\times C}$. A single reference leaves a smooth manifold $\mathcal{S}$ whose dimension depends on the regime. Several references, under simultaneous-diagonalisation conditions on the reference covariances, reduce the matching set $\mathcal{S}_K$, everywhere on $\mathbb{R}^{P\times C}$, to $2^{\min(P,C)}$ isolated solutions, mutually Frobenius-separated, which differ from one another only by sign flips of the rows in the whitened coordinates of Σ^{ref}. The gradient bound tells us flow through $A^\dagger$ stays controlled while UV^T is small relative to $\sigma_{\min}(A_0)$; the controlled scale tightens as rank grows, so rank has a usable ceiling. The trajectory result tells us that, starting from A_0, gradient flow stays nearest the same solution as $B(A_0) := \Pi A_0^\dagger$. The other solutions are loss-equivalent and Frobenius-far from $B(A_0)$, so confinement keeps A a small correction to A_0 throughout training, consistent with the LoRA premise (low-rank adaptation of pre-trained networks, Hu et al. (2022)).

1 Introduction

Combining measurements of a shared hidden source from multiple modalities is a recurrent task in inverse problems. Direct sample-level matching across modalities is typically unavailable. The modalities may be sampled on incompatible time bases, carry unknown relative delays, or lack a shared spatial event grid. Often the cross-modal relation has to be expressed at a coarser level than individual samples. Second-order statistics survive the alignment loss and have long been used to combine modal evidence in inverse problems (Cardoso & Souloumiac, 1996; Belouchrani et al., 1997; Ribot et al., 2025). The present paper is concerned with the question of identifying the linear operator from the covariances (Belouchrani et al., 1997; Aissa El Bey et al., 2008; Vía et al., 2011). For example, in an EEG–fMRI experiment, a subject performs a task while neural electrical activity is observed in two modalities: at the scalp via EEG, and via the haemodynamic response sampled by fMRI. A generic head model supplies an approximate cortex-to-scalp-EEG forward operator. The subject-specific operator is the unknown to be recovered from the paired second-order statistics of the two modalities. Here Π maps cortical sources onto the atlas parcels on which the fMRI signal is read out (Fischl et al., 2004), and A_0 is the lead field of a boundary-element head model (Mosher et al., 1999).

We formalize the setting as a linear forward model $y \approx Ax$, $A \in \mathbb{R}^{C\times N}$, $C < N$. The task is to recover the unknown operator A from paired second-order statistics (i) Σ_y of y and (ii) Σ^{ref} of *aggregated* Πx, where $\Pi \in \mathbb{R}^{P\times N}$ is a known aggregator with $P < N$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

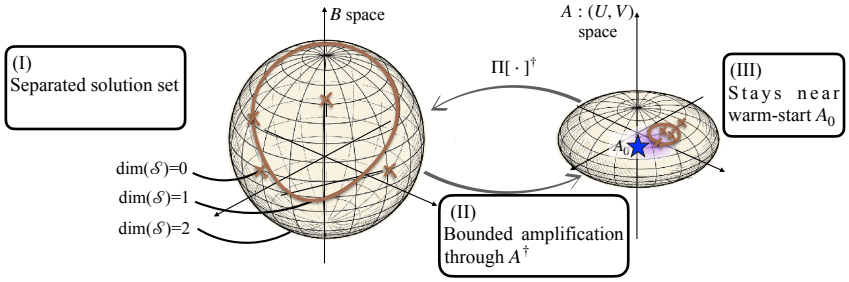


Figure 1: Illustration of main objects and results of the paper. Matching is through B-space (left), where the structure of the solution set $\mathcal{S}$ is discussed. The second result discusses the non-linear relationship to A (right) due to the pseudo-inverse. The third result establishes desirable training behaviour when warm-starting from a near-optimal A_0 (training does not cause jumping between B-classes).

We parameterize the recovery of A as a low-rank correction of the known baseline A_0,

$$A \;=\; A_0 \;+\; UV^T, \qquad U \in \mathbb{R}^{C\times r},\; V \in \mathbb{R}^{N\times r},\; r < \min(C, N), \tag{1}$$

where r is the rank of the correction. In this model, the covariance-matching objective is represented via Moore–Penrose pseudo-inversion using $\hat{\Sigma} := B\Sigma_y B^T$ with $B := \Pi A^\dagger$, representing recovery-then-aggregation under A. The objective is to match $\hat{\Sigma}$ to Σ^{ref}.

We address three aspects of training under this objective; Figure 1 summarises them geometrically.[1] In short, the covariances fix B up to sign flips, and training from A_0 keeps the correct one. **(I)** *Solution-space structure of B.* We characterise the matching set: when the problem pins B down to a discrete set of solutions, how many such solutions there are, and how far apart they sit. **(II)** *Pseudo-inverse amplification.* We bound the worst-case Frobenius amplification of $dA^\dagger$, the differential of the map $A \mapsto A^\dagger$, in terms of $\sigma_{\min}(A)$, and show it grows monotonically with the rank of the correction UV^T — giving a usable rank ceiling for training. **(III)** *Training stays near A_0.* Combining (I) and (II), gradient flow from A_0 stays close to the warm-start solution throughout training: the LoRA premise — that A remains a small correction to A_0 — holds dynamically, not only at initialisation.

2 Related work

The matching set $\mathcal{S}_K$ of §4.1, defined after (5), is, when the references are exactly attainable, the joint zero set in B of the K equations $B\Sigma_y^{(k)}B^T = \Sigma^{\text{ref},(k)}$, the structural object of joint blind source separation. Identifiability for joint diagonalisation of K symmetric matrices is established by Vía et al. (2011) and surveyed by Pan et al. (2022); Aissa El Bey et al. (2008) treat the SOBI parameterisation (Belouchrani et al., 1997) and prove identifiability up to a $2^P P!$ sign-permutation indeterminacy under pairwise non-proportionality of source autocorrelations. The classical algorithmic line on simultaneous diagonalisation of covariance pencils is Cardoso & Souloumiac (1996); Belouchrani et al. (1997); Pham (2001). Proposition 3 specialises Aissa El Bey et al. (2008)'s Theorem 1 to the real case for $P \geq C$, and extends it to $P < C$ under the further hypotheses (6) and (7). Proposition 4 adds the Frobenius gap δ_B between sign-classes, absent from this literature.

The closed-form differential $dA^\dagger[\Delta A]$ in (10) is classical: Golub & Pereyra (1973) establish it for variable-projection nonlinear least squares, and Stewart (1969); Wedin (1972; 1973) develop the perturbation theory of the pseudo-inverse and the SVD; we use the form of Papadopoulo & Lourakis (2000). Algorithmic alternatives for back-propagation through pseudo-inverses, SVD, and matrix square roots include Newton–Schulz iteration (Song et al., 2022), truncated-Taylor SVD differentiation (Wang et al., 2022; Zhang & Hu, 2024; Wang et al., 2025; Krämer et al., 2024), and condition-number penalisation (Nenov et al., 2024); these control inner-iteration accuracy rather than back-prop amplification. We use the closed form to extract the Frobenius Lipschitz constant $2/\sigma_{\min}(A)^2$ (Lemma 5) and the rank-monotone bound $\bar{\kappa}(k)$ (Proposition 6).

[1] Figure prepared as follows: geometric layout computed programmatically; pen-and-ink stylisation by an LLM.

The parameterisation $A = A_0 + UV^T$ (Equation 1) coincides in shape with low-rank adaptation (Hu et al., 2022), which targets parameter-efficient fine-tuning under full supervision; we study identifiability and gradient-flow stability of the same shape under a covariance-matching loss with no pointwise supervision of x. The pseudo-inverse step (ii) of Equation 4, central to our amplification bound, is absent from the adaptation setting.

Forward-operator learning treats the forward map as unknown or partially known and learns it from data; see Ongie et al. (2020); Arridge et al. (2024) for reviews and Adler & Öktem (2017) for a representative method. This line predominantly uses a fully data-driven operator (no A_0) with paired (x, y) supervision or generative priors on x; ours fixes A_0 as a warm-start and supervises through the second-order statistic of Πx. Proposition 7 couples these into a quantitative trajectory bound, with no analogue we are aware of in this literature.

A separate asymptotic line analyses the distribution of joint-diagonalisation estimators as $T \to \infty$ at fixed identifiability gap (Tichavský & Yeredor, 2009; Ablin et al., 2019; Ribot et al., 2025); Proposition 4 concerns the non-asymptotic population gap and does not overlap.

3 Setup and notation

Operators	*Covariances and loss*
$A \in \mathbb{R}^{C\times N}$, full row rank	$\Sigma_y \in \mathrm{Sym}_C^+$, channel covariance
A_0 known baseline; $A = A_0 + UV^T$	$\Sigma^{\mathrm{ref}} \in \mathrm{Sym}_P^+$, reference (given)
U, V $U \in \mathbb{R}^{C\times r}$, $V \in \mathbb{R}^{N\times r}$	$\widehat{\Sigma} := B\Sigma_y B^T$, predicted reference
r rank of UV^T (see $r_{\max}$, Eq. (12))	$\mathcal{L}(\cdot)$ training loss, depends on A, B or (U, V)
ρ equal column norm, $\|u_i\|_2 = \|v_i\|_2 = \rho$	$\mathcal{S} := \arg\min_B \mathcal{L}$, matching set
$\Pi \in \mathbb{R}^{P\times N}$, fixed aggregator	$\mathcal{S}_K := \arg\min_B \mathcal{L}_K$, see (5)
$B := \Pi A^\dagger \in \mathbb{R}^{P\times C}$	λ_i i-th largest eigenvalue of Σ^{ref}
$B(A_0) := \Pi A_0^\dagger$, value of B at A_0	$d[\cdot]$ $d[\cdot][\Delta]$, linear-in-Δ part

We briefly introduce the framework and model setup, and fix commonly used notation as in the table above. The forward model and covariance-matching have been described in the introduction:

$$y \approx Ax, \qquad \widehat{\Sigma} = B\Sigma_y B^T, \tag{2}$$

where we recall the model for A in (1), $A = A_0 + UV^T$. The training objective is

$$\mathcal{L} = \left\|\widehat{\Sigma} - \Sigma^{\mathrm{ref}}\right\|_F^2. \tag{3}$$

The dependency chain from the parameters (U, V) to the loss $\mathcal{L}$ in this learning framework is

$$(U,V) \xrightarrow[\text{(i) bilinear}]{A_0+UV^T} A \xrightarrow[\text{(ii) inversion}]{A^\dagger} A^\dagger \xrightarrow[\text{(iii) linear}]{\Pi\,\cdot} B \xrightarrow[\text{(iv) quad}]{\cdot\,\Sigma_y\,\cdot^T} \widehat{\Sigma} \xrightarrow[\text{(v) quad}]{\|\cdot-\Sigma^{\mathrm{ref}}\|_F^2} \mathcal{L}. \tag{4}$$

First note that there are certain classes of reparameterizations that leave $\mathcal{L}$ unchanged, which are NOT the focus of this paper. We will fix them by convention. In step (i), $(U, V) \mapsto A_0 + UV^T$ admits an r-dimensional family of rotational symmetries: $(U, V) \sim (UR, R^{-T}V)$ for any $R \in \mathrm{GL}_r$. We will constrain U and V to have equal column norms, $\|u_i\|_2 = \|v_i\|_2 = \rho$, $i = 1, \ldots, r$ by reprojection during training. The sign-flip symmetry of u_i and v_i is irrelevant. In step (iii), Π is a fixed linear contraction $\mathbb{R}^N \to \mathbb{R}^P$, $P < N$; non-invertible in the forward, trivial Π^T in the backward.

The remainder of the analysis walks the chain in reverse, from the loss back to the parameters. §4.1 characterises the matching set $\mathcal{S}$ in B-space — its dimension and component structure, in single- and multi-covariance form. §4.2 differentiates step (ii), $A \mapsto A^\dagger$, and uses this differential to pull $\mathcal{S}$ back to A-space; the rank r of the correction enters explicitly here. §4.3 reaches (U, V) and studies the gradient flow when a solution lies near the baseline A_0.

4 Analysis

4.1 The matching set in B-space

Define

$$\varepsilon^* := \min_{B\in\mathbb{R}^{P\times C}} \left\|B\Sigma_y B^T - \Sigma^{\text{ref}}\right\|_F^2, \qquad \mathcal{S} := \arg\min_B \left\|B\Sigma_y B^T - \Sigma^{\text{ref}}\right\|_F^2.$$

When the reference is exactly attainable, $\varepsilon^* = 0$ and $\mathcal{S}$ is the zero set of the matching equation $B\Sigma_y B^T = \Sigma^{\text{ref}}$; when rank obstruction binds, $\varepsilon^* > 0$ and $\mathcal{S}$ is the set of best feasible fits.

Proposition 1. *$\mathcal{S}$ is a compact smooth submanifold of $\mathbb{R}^{P\times C}$:* Underdetermined regime, $P \le C$. *Matching is exact, $\varepsilon^* = 0$. For any Cholesky factor R of Σ_y, the map $B \mapsto (\Sigma^{\text{ref}})^{-1/2} B\, R$ is a diffeomorphism from $\mathcal{S}$ onto the Stiefel manifold $V_P(\mathbb{R}^C) := \{O \in \mathbb{R}^{P\times C} : OO^T = I_P\}$.* $\dim \mathcal{S} = PC - P(P{+}1)/2$.

Overdetermined regime, $P > C$. *The image $\{B\Sigma_y B^T : B \in \mathbb{R}^{P\times C}\}$ is the rank-$\le C$ p.s.d. cone, so by Eckart–Young $\varepsilon^* = \sum_{i=C+1}^{P} \lambda_i(\Sigma^{\text{ref}})^2 > 0$, where $\lambda_1(\Sigma^{\text{ref}}) \ge \cdots \ge \lambda_P(\Sigma^{\text{ref}})$ are the eigenvalues of Σ^{ref} in decreasing order, and $\mathcal{S}$ is a single $O(C)$-orbit, of dimension $C(C{-}1)/2$.*

The differential geometry of the Stiefel manifold is standard; see, e.g., Edelman et al. (1998).

A single reference leaves $\mathcal{S}$ continuously degenerate (positive-dimensional unless $\min(P, C) = 1$). To tighten this, it is common to collect multiple (K) reference covariances of the same source, e.g. from distinct time-lags or experimental conditions. Let $\Sigma_y^{(k)}, \Sigma^{\text{ref},(k)}$ be the k-th pair, $\widehat{\Sigma}^{(k)} := B\Sigma_y^{(k)} B^T$, and the multi-covariance loss

$$\mathcal{L}_K(A) \;:=\; \sum_{k=1}^{K} w_k \left\|\widehat{\Sigma}^{(k)} - \Sigma^{\text{ref},(k)}\right\|_F^2, \qquad w_k > 0, \tag{5}$$

where we let $\varepsilon_K^* := \min_B \mathcal{L}_K(B)$ and $\mathcal{S}_K := \arg\min_B \mathcal{L}_K(B)$ be the counterparts of ε^* and $\mathcal{S}$ for the multi-covariance setting. Note that a signal can supply *cross-covariance* between the signal and lag-τ version of itself $\mathbb{E}[\Pi x_t (\Pi x_{t-\tau})^T]$ in the multi-reference setting. We will refer to this as a *non-symmetric* reference.

The identifiability question arises from the lack of supervision. The remainder of this subsection therefore treats the structure of the solution set $\mathcal{S}$ and $\mathcal{S}_K$ in the under-determined regime $P \le C$, in which they are largest. Our final result on the training and optimisation of the entire framework remains valid in both regimes, which will be discussed later in §4.3. The following results are stated for $\varepsilon_* = \varepsilon_K^* = 0$, i.e. the reference(s) are simultaneously attainable by some B. Appendix A.3 gives a sufficient condition on the data under which this holds.

Perturb $B \to B + \Delta B$ in the k-th predicted reference $B\Sigma_y^{(k)} B^T$. The first-order change is

$$(B+\Delta B)\Sigma_y^{(k)}(B+\Delta B)^T - B\Sigma_y^{(k)} B^T \;=\; L_k(\Delta B) \;+\; (\Delta B)\Sigma_y^{(k)}(\Delta B)^T,$$

where $L_k(\Delta B) := (\Delta B)\Sigma_y^{(k)} B^T + B\Sigma_y^{(k)}(\Delta B)^T$ is the linear-in-ΔB part. We use $d[\cdot][\Delta]$ throughout for the linear-in-Δ part of a perturbation: $L_k(\Delta B) = d[B\Sigma_y^{(k)} B^T][\Delta B]$. Stack these into $L_B := (L_1, \ldots, L_K) : \mathbb{R}^{P\times C} \to \bigoplus_k \mathbb{R}^{n_k}$, where $n_k = P(P{+}1)/2$ for symmetric references and P^2 for non-symmetric ones.

At $B \in \mathcal{S}_K$ with $\varepsilon_K^* = 0$, a direction $\Delta B \in \ker L_B$ is one along which all K matches are preserved to first order; it is tangent to $\mathcal{S}_K$. The solution set is locally a single point near B if $\ker L_B = \{0\}$, equivalently $\operatorname{rank} L_B = PC$. We call this the *rank condition*. The converse fails when $P < C$ (Appendix A.2).

Proposition 2. *Let $B_\star \in \mathcal{S}_K$. The rank condition* $\operatorname{rank} L_{B_\star} = PC$ *implies the counting condition $\sum_k n_k \ge PC$, and it is sufficient for $\mathcal{S}_K$ to be discrete near $B_\star$.*

The implication follows from the dimension count $\operatorname{rank} L_{B_\star} \le \sum_k n_k$. Discreteness follows from the implicit-function theorem at $B_\star$. See Appendix A for the full proof.

A non-symmetric reference ($n_k = P^2$) supplies more constraints than a symmetric one ($n_k = P(P{+}1)/2$); L_B re-appears as the kernel of the multi-covariance Hessian (Theorem 8).

Specialization under simultaneous diagonalisation. The base pair $(\Sigma_y, \Sigma^{\text{ref}})$ is one of the K pairs, so $\mathcal{S}_K$ is the zero set of all the equations quantified over below. Diagonalize $\Sigma^{\text{ref}} = Q\Lambda Q^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_P)$. Write $\Psi := \Lambda^{-1/2} Q^T$ for the whitening transform of Σ^{ref}; when $P > C$ it is taken on the rank-C part of Σ^{ref} (Appendix A.1). Define $\widetilde{B} := \Psi B$ and $D_k := \Lambda^{-1/2} Q^T \Sigma^{\text{ref},(k)} Q \Lambda^{-1/2}$, so the matching conditions read $\widetilde{B}\Sigma_y \widetilde{B}^T = I_P$ and $\widetilde{B}\Sigma_y^{(k)} \widetilde{B}^T = D_k$.

Proposition 3. *Suppose $\lambda_1 > \cdots > \lambda_{\min(P,C)} > 0$, each D_k is diagonal, and the row vectors $\boldsymbol{\rho}_i := (d_{i1}, \ldots, d_{iK}) \in \mathbb{R}^K$ ($i = 1, \ldots, \min(P, C)$) are pairwise distinct. When $P < C$, suppose in addition that (6) and (7), stated immediately after this proposition, hold. Then any B with $B\Sigma_y B^T = \Sigma^{\text{ref}}$ and $B\Sigma_y^{(k)} B^T = \Sigma^{\text{ref},(k)}$ for all k is determined up to the sign indeterminacy of order $2^{\min(P,C)}$ acting in the whitened coordinates by $\widetilde{B} \mapsto D_\pm \widetilde{B}$, that is by $B \mapsto \Psi^{-1} D_\pm \Psi B = Q\Lambda^{1/2} D_\pm \Lambda^{-1/2} Q^T B$, with $D_\pm \in \{\pm 1\}^{\min(P,C)}$ diagonal, padded with $+1$ on the remaining $P - C$ coordinates when $C < P$.*

The two further hypotheses are stated in terms of the whitened channel covariances $G_k := \Sigma_y^{-1/2} \Sigma_y^{(k)} \Sigma_y^{-1/2}$:

$$G_k = \Omega \, \text{diag}(\gamma_{1k}, \ldots, \gamma_{Ck}) \, \Omega^T, \quad k = 1, \ldots, K, \text{ for some } \Omega \in O(C); \tag{6}$$

$$\boldsymbol{\rho}_i \text{ is a vertex of } \text{conv}\{\gamma_1, \ldots, \gamma_C\} \text{ and equals } \gamma_c \text{ for exactly one } c, \quad i = 1, \ldots, P, \tag{7}$$

where $\gamma_c := (\gamma_{c1}, \ldots, \gamma_{cK}) \in \mathbb{R}^K$ collects the c-th whitened channel response across the K conditions, and that index c is written $c(i)$. The role of both conditions is discussed in Appendix A.1.

The exponent $\min(P, C)$ accounts for $C < P$: $\widetilde{B}$ is $P \times C$ with at most C independent rows. For $C \geq P$ this recovers 2^P. Prescribing the diagonal removes the $P!$ permutation factor present in the standard SOBI identifiability result, where the diagonal ordering is not fixed (Aissa El Bey et al., 2008). The proof is in Appendix A.1.

Proposition 4. *Adopt the hypotheses of Proposition 3. For any two distinct $B_0, B_1 \in \mathcal{S}_K$,*

$$\|B_0 - B_1\|_F \;\geq\; \frac{2\,\lambda_i^{1/2}}{\sigma_{\max}(\Sigma_y)^{1/2}} \tag{8}$$

for some $i \in \{1, \ldots, \min(P, C)\}$. Define $\delta_B := 2\,\lambda_{\min(P,C)}^{1/2} / \sigma_{\max}(\Sigma_y)^{1/2}$.

Sketch. Pass to whitened coordinates as above, where the sign indeterminacy of Proposition 3 acts as $\widetilde{B} \mapsto D_\pm \widetilde{B}$. From $\widetilde{B}\Sigma_y \widetilde{B}^T = I_P$ and the i-th diagonal entry, $\widetilde{b}_i^T \Sigma_y \widetilde{b}_i = 1$; sub-multiplicativity gives $\|\widetilde{b}_i\|_2^2 \geq 1/\sigma_{\max}(\Sigma_y)$. A sign-flip on row i alone changes $\widetilde{B}$ by $2e_i e_i^T \widetilde{B}$, contributing $4\|\widetilde{b}_i\|_2^2$ to $\|\widetilde{B}_0 - \widetilde{B}_1\|_F^2$. Un-whitening multiplies row i by $\lambda_i^{1/2}$, giving $\|B_0 - B_1\|_F^2 \geq 4\lambda_i / \sigma_{\max}(\Sigma_y)$. Multi-row sign-flips only enlarge the sum. □

The right-hand side of (8) depends only on Σ^{ref} and $\sigma_{\max}(\Sigma_y)$, both spectral invariants computable before optimisation.

4.2 The pseudo-inverse differential and rank dependence

Lemma 5. *For $A \in \mathbb{R}^{C \times N}$ of full row rank and $\Delta A \in \mathbb{R}^{C \times N}$,*

$$\|dA^\dagger[\Delta A]\|_F \;\leq\; \frac{2}{\sigma_{\min}(A)^2} \, \|\Delta A\|_F. \tag{9}$$

Proof. The pseudo-inverse differential, in the $d[\cdot][\Delta]$ convention introduced in §4.1, is classical (Golub & Pereyra, 1973; Papadopoulo & Lourakis, 2000):

$$dA^\dagger[\Delta A] \;=\; -\,A^\dagger\,(\Delta A)\,A^\dagger \;+\; (I_N - A^\dagger A)\,(\Delta A)^T\,(AA^T)^{-1}, \tag{10}$$

proved by differentiating $G \cdot (AA^T) = I_C$ for $G := (AA^T)^{-1}$. Bound each term of (10) using $\|XYZ\|_F \leq \|X\|_{\text{op}} \|Y\|_F \|Z\|_{\text{op}}$. For the first term, $\|A^\dagger\|_{\text{op}} = \sigma_{\min}(A)^{-1}$ gives $\|A^\dagger(\Delta A)A^\dagger\|_F \leq \sigma_{\min}(A)^{-2} \|\Delta A\|_F$. For the second, $\|I_N - A^\dagger A\|_{\text{op}} = 1$ and $\|(AA^T)^{-1}\|_{\text{op}} = \sigma_{\min}(A)^{-2}$ give the same bound. Adding the two contributions yields (9). □

The closed-form differential and the explicit Frobenius bound (9) avoid the inner convergence parameter of Newton–Schulz (Song et al., 2022), the eigenvalue-gap threshold of truncated-Taylor SVD differentiation (Wang et al., 2022; Zhang & Hu, 2024), and the implicit dynamics control of condition-number penalisation (Nenov et al., 2024). The bound supplies a forward Lipschitz constant $2\sigma_{\min}(A)^{-2}$ used in the rank constraint below.

Fix a design target $\kappa_0 > 0$ with $\sigma_{\min}(A_0) > \sqrt{2/\kappa_0}$. Weyl's inequality $\sigma_{\min}(A) \geq \sigma_{\min}(A_0) - \|UV^T\|_{\text{op}}$ gives $2/\sigma_{\min}(A)^2 \leq \kappa_0$ whenever

$$\|UV^T\|_{\text{op}} \;\leq\; \sigma_{\min}(A_0) \;-\; \sqrt{2/\kappa_0}. \tag{11}$$

We write $\mathcal{F}_A := \{A = A_0 + UV^T : 2/\sigma_{\min}(A)^2 \leq \kappa_0\}$.

We constrain (U, V) throughout to per-column equal norms $\|u_i\|_2 = \|v_i\|_2 = \rho$ for $i = 1, \ldots, r$. The equal-norm-at-each-i part places no restriction on UV^T: for any (U, V) producing a given UV^T, rescaling each column pair $(u_i, v_i) \to (\alpha_i u_i,\, v_i/\alpha_i)$ preserves $u_i v_i^T$ (hence UV^T); choosing $\alpha_i = \sqrt{\|v_i\|_2/\|u_i\|_2}$ equalises $\|u_i\|_2$ and $\|v_i\|_2$ at each i. The further assumption that ρ is shared across i is a real constraint on UV^T, maintained in training by per-step renormalisation (see Appendix D; see also §4.3.1 for the relation to the gradient-flow analysis of §4.3).

Proposition 6. *Under this normalisation, $\|UV^T\|_{\text{op}} \leq r\rho^2$, and $\bar{\kappa}(r) := \frac{2}{(\sigma_{\min}(A_0) - r\rho^2)^2}$ is strictly increasing in r on $\{r\rho^2 < \sigma_{\min}(A_0)\}$.*

Proof. $UV^T = \sum_i u_i v_i^T$, so by the triangle and sub-multiplicative inequalities $\|UV^T\|_{\text{op}} \leq \sum_i \|u_i\|_2 \|v_i\|_2 = r\rho^2$. Substituting $\sigma_{\min}(A) \geq \sigma_{\min}(A_0) - r\rho^2$ into Lemma 5 produces $\bar{\kappa}(r)$. □

Combining (11) with $\|UV^T\|_{\text{op}} \leq r\rho^2$ from Proposition 6 gives the upper bound on usable rank

$$r_{\max} \;:=\; \left\lfloor \frac{\sigma_{\min}(A_0) - \sqrt{2/\kappa_0}}{\rho^2} \right\rfloor. \tag{12}$$

The large-r end is ruled out by admissibility: ranks above $r_{\max}$ violate (11), and within the admissible range Proposition 6 gives $\bar{\kappa}(r)$ strictly increasing. The small-r end is empirical: rank one is too rigid representationally.

Trajectory bound under warm start. In the following discussion, we will loosely use the notation $B(A)$ or $B(U, V)$ to denote the dependence of B on the arguments, as illustrated in the chain diagram (4).

Lemma 5 bounds $dA^\dagger$ at fixed A; integrating along $A_0 \to A_0 + UV^T$ yields a bound on $\|B(U, V) - B(A_0)\|_F$.

Proposition 7. *Fix $A_0 \in \mathbb{R}^{C\times N}$ of full row rank, with $\sigma_0 := \sigma_{\min}(A_0)$, $\Pi \in \mathbb{R}^{P\times N}$, $B(A) := \Pi A^\dagger$, and r the training rank of UV^T. Set*

$$c \;:=\; \frac{8\sqrt{r}\,\|\Pi\|_{\text{op}}}{\sigma_0^2}. \tag{13}$$

(1) *Call (U, V)* admissible *if $\|UV^T\|_{\text{op}} \leq \rho \leq \sigma_0/2$ for some $\rho > 0$; this is the assumption used throughout the present proposition and below. For every admissible (U, V),*

$$\big\|B(A_0 + UV^T) \;-\; B(A_0)\big\|_F \;\leq\; c\,\rho. \tag{14}$$

This bound requires no hypothesis on ε_K^* or on $\mathcal{S}_K$.

(2) *Suppose $\varepsilon_K^* = 0$, that there is a unique $B_\star \in \mathcal{S}_K$ minimising $\|B(A_0) - B_\star\|_F$, and*

$$c\,\rho \;+\; \|B(A_0) - B_\star\|_F \;<\; \tfrac{1}{2}\,\delta_B. \tag{15}$$

Then for every admissible (U, V), $B(A_0 + UV^T)$ lies in $\mathcal{B}(B_\star, \delta_B/2)$, the open Frobenius ball of radius $\delta_B/2$ about $B_\star$, and $B_\star$ is the unique nearest element of $\mathcal{S}_K$ to $B(A_0 + UV^T)$.

(3) *Under the same hypothesis, the trajectory $\{B_t\}_{t\geq 0} := \{B(A_0 + U_t V_t^T)\}_{t\geq 0}$ of gradient flow inside the admissible region satisfies, for all admissible t,*

$$\big\|\text{Proj}_{\ker L_{B_\star}}(B_t - B_\star)\big\|_F \;\leq\; D_\|, \qquad D_\| \;:=\; c\,\rho \;+\; \|B(A_0) - B_\star\|_F. \tag{16}$$

$D_\parallel$ is an a-priori upper bound on the $\ker L_{B_\star}$-component of $B_t - B_\star$ throughout admissible training; it re-appears in the sublevel-set splitting of Lemma 9 as the slack on $\|\Delta_\parallel\|_F$.

Sketch. (1) Apply Lemma 5 along the segment $A(s) := A_0 + sUV^T$, $s \in [0,1]$. Weyl gives $\sigma_{\min}(A(s)) \geq \sigma_0 - \rho \geq \sigma_0/2$, hence $2/\sigma_{\min}(A(s))^2 \leq 8/\sigma_0^2$. Integrating and using $\|UV^T\|_F \leq \sqrt{r}\,\|UV^T\|_{\mathrm{op}} \leq \sqrt{r}\,\rho$ for $\mathrm{rank}(UV^T) \leq r$, then composing with Π, gives $c\rho$ with the stated c. (2) Triangle inequality with (14) and (15) places $B(A_0 + UV^T)$ within $\delta_B/2$ of $B_\star$; Proposition 4 separates other sign-classes by δ_B. (3) Frobenius-projection non-expansiveness on $\|B_t - B_\star\|_F \leq c\rho + \|B(A_0) - B_\star\|_F$. See Appendix B for the full proof. □

Proposition 7 is a consequence of Lemma 5, Weyl's inequality, Proposition 4, and the triangle inequality, and does not constitute a result inaccessible to prior art. Its role is structural: in the $C > P$ regime ($\dim\ker L_{B_\star} > 0$), part (3) supplies the a-priori upper bound on $\|\mathrm{Proj}_{\ker L_{B_\star}}(B - B_\star)\|_F$ that Lemma 9 requires; without (3) the floor on $\|\mathrm{Proj}_{\ker L_{B_\star}^\perp}(B - B_\star)\|_F^2$ in that argument does not close. In the $C \leq P$ regime, the parallel component vanishes, so $D_\parallel$ may be replaced by 0 in (18), and Lemma 9 reduces to its standalone form.

4.3 Gradient flow near A_0

There are a few conditions that the results in this subsection require: (a) the simultaneous-diagonalisation condition of Proposition 3, together with (6) and (7) when $P < C$; (b) the Hessian (17) satisfies $H[\Delta] \geq h\|\Delta\|_F^2$ on $\ker L_{B_\star}^\perp$ with h above the threshold (28) of Appendix C, which gives the trust-region radius $r_0 \geq \delta_B/2$ used by Lemma 9. When $\varepsilon_K^* > 0$, the residual cross-term $2\langle \Gamma_0^{(k)}, \Delta\Sigma_y^{(k)}\Delta^T\rangle$ in (17) is folded into the verification of h; (c) at some finite t_0 the trajectory satisfies $B_{t_0} \in \mathcal{B}(B_\star, \delta_B/2)$ and $\mathcal{L}_K(B_{t_0}) - \varepsilon_K^* < \alpha^\star$, with $\alpha^\star$ from Lemma 9. The full list is in Appendix C.

Theorem 8. *At any $B_\star \in \arg\min \mathcal{L}_K$, the Hessian of $\mathcal{L}_K$ acting on $\Delta \in \mathbb{R}^{P\times C}$, by which we mean the second-order term $H[\Delta]$ of the expansion of $\mathcal{L}_K$ at $B_\star$, equal to half the second directional derivative along Δ, equals*

$$H[\Delta] \;=\; \sum_{k=1}^{K} w_k \Big(\|L_k(\Delta)\|_F^2 \;+\; 2\,\langle \Gamma_0^{(k)},\, \Delta\Sigma_y^{(k)}\Delta^T\rangle_F \Big), \tag{17}$$

where $L_k(\Delta) := \Delta\Sigma_y^{(k)}B_\star^T + B_\star\Sigma_y^{(k)}\Delta^T$ and $\Gamma_0^{(k)} := B_\star\Sigma_y^{(k)}B_\star^T - \Sigma^{\mathrm{ref},(k)}$. When $\varepsilon_K^* = 0$, $\Gamma_0^{(k)} = 0$ for every k, (17) reduces to $\sum_k w_k\|L_k(\Delta)\|_F^2$, and rank H = rank L *for $L = (L_1, \ldots, L_K)$.*

Sketch. Expand $\widehat{\Sigma}^{(k)}(B_\star + \Delta) - \Sigma^{\mathrm{ref},(k)} = \Gamma_0^{(k)} + L_k(\Delta) + \Delta\Sigma_y^{(k)}\Delta^T$. Squaring, summing with weights w_k, and grouping by order in Δ, the linear-in-Δ contribution is $\sum_k w_k\langle\Gamma_0^{(k)}, L_k(\Delta)\rangle_F$, which vanishes by stationarity at $B_\star$; the quadratic-in-Δ contribution gathers into (17); the remainder is cubic and quartic in Δ. Full proof in Appendix C. □

When $\varepsilon_K^* = 0$, the kernel of H at $B_\star$ coincides with $\ker L_{B_\star} := \bigcap_k \ker L_k$. It is trivial exactly when the rank condition holds, and it can be strictly larger than the tangent space of $\mathcal{S}_K$ (Appendix A.2). When $\varepsilon_K^* > 0$, the residual cross-term in (17) can shift the kernel; condition (b) of Appendix C absorbs both regimes by requiring $H \succeq hI$ on $\ker L_{B_\star}^\perp$ with $h > 0$.

Lemma 9. *Fix $B_\star \in \mathcal{S}_K$ with h and δ_B. Suppose conditions (a) and (b) of Appendix C, and suppose that $D_\parallel \geq 0$ is such that the B considered below satisfy $\|\mathrm{Proj}_{\ker L_{B_\star}}(B - B_\star)\|_F \leq D_\parallel$. Set*

$$\alpha^\star \;:=\; \tfrac{h\,\delta_B^2}{8} \;-\; \tfrac{2\,h\,D_\parallel^2}{3}. \tag{18}$$

If $D_\parallel < \delta_B/\sqrt{12}$, which ensures $\alpha^\star > 0$, then

$$\big\{B \in \mathcal{B}(B_\star, \delta_B/2) : \mathcal{L}_K(B) - \varepsilon_K^* < \alpha^\star\big\} \;\subset\; \big\{B : \|\mathrm{Proj}_{\ker L_{B_\star}^\perp}(B - B_\star)\|_F \;<\; \delta_B/2\big\},$$

and the left-hand set is an open neighbourhood of $B_\star$ containing no other element of $\mathcal{S}_K$. Moreover every B with $\|B - B_\star\|_F = \delta_B/2$ satisfies $\mathcal{L}_K(B) - \varepsilon_K^* \geq \alpha^\star$. ($\alpha^\star$ is an excess-loss threshold; under $\varepsilon_K^* = 0$ the constraint reduces to $\mathcal{L}_K(B) < \alpha^\star$.)

In the $C \leq P$ regime, $\ker L_{B_\star} = 0$ and $D_\parallel = 0$ is admissible, so $\alpha^\star = h\delta_B^2/8$ and the warm-start admissibility hypothesis is no longer required.

The proof is in Appendix C.

Corollary 10. *Let A_t be a gradient-flow trajectory of $\mathcal{L}(A) := \mathcal{L}_K(\Pi A^\dagger)$ starting in $\mathcal{F}_A$, $B_t := \Pi A_t^\dagger$, and $B_\star \in \mathcal{S}_K$ the unique nearest element to B_{t_0}. Under conditions (a)–(c) of Appendix C, for all $t \geq t_0$, B_t lies in $\mathcal{B}(B_\star, \delta_B/2)$.*

The proof is in Appendix C.

Main result. Propositions 3 and 4 show that, over the whole of $\mathbb{R}^{P\times C}$, the matching set $\mathcal{S}_K$ consists of the $2^{\min(P,C)}$ sign-flip images of one solution, pairwise at least δ_B apart. Under (15), Proposition 7 keeps every admissible $B(A_0 + UV^T)$ within $\delta_B/2$ of the solution $B_\star$ nearest $B(A_0)$, and Corollary 10 keeps gradient flow there at all later times. Together they give the result stated in §1: the covariances fix B up to sign flips, and training from A_0 keeps the correct one.

4.3.1 Per-step renormalisation and the gradient-flow analysis

Corollary 10 is stated for continuous gradient flow on A. Practical algorithms run discrete gradient steps on (U, V) followed by a per-step equal-column-norm projection. The projection picks an arbitrary member from an equivalence class (GL_r gauge representative) at each step but stays within the same fibre of A-space; hence the discrete trajectory in A-space $A_t = A_0 + U_t V_t^T$ is well-defined under projection. The relation between this discrete trajectory and the continuous gradient flow on A is as follows.

For $\Delta \in \ker L_{B_\star}$, the change in the predicted observable is Δy with second moment $\mathbb{E}[\|\Delta y\|^2] = \mathrm{tr}(\Delta \Sigma_y \Delta^T) \leq \sigma_{\max}(\Sigma_y)\, \|\Delta\|_F^2$, and $\|\Delta\|_F \leq D_\parallel = c\rho + \|B(A_0) - B_\star\|_F$ by Proposition 7 (3). The matching loss is invariant to first order under Δ; the sample-level prediction is not, with the change bounded by $\sigma_{\max}(\Sigma_y)\, D_\parallel^2$.

Remark 11. Let $\mathcal{F}_A^{(0)}$ be the path-connected component of $\mathcal{F}_A$ containing A_0. On $\mathcal{F}_A^{(0)}$, $A \mapsto \Pi A^\dagger$ is uniformly Lipschitz with constant $\kappa_0 \|\Pi\|_{\mathrm{op}}$ (Lemma 5). For any trajectory $A_t \in \mathcal{F}_A^{(0)}$ with

$$\sum_t \|\Delta A_t\|_F \;<\; \frac{\delta_A}{2} \;:=\; \frac{\delta_B}{2\,\kappa_0\,\|\Pi\|_{\mathrm{op}}}, \tag{19}$$

$\Pi A_T^\dagger$ and $\Pi A_0^\dagger$ are in the same sign-class of $\mathcal{S}_K$ under the action of Proposition 3. This alternative uses only the first-order amplification information of §4.2; the trade-off is that the bound is a path-length integral rather than a sublevel-set inclusion, and assumption (c) is not used.

Numerical illustration. On a small calibration instance ($N = 28^2$, $C = 64$, Π a 4×4 average-pooling map giving $P = 49$, planted rank-4 multiplicative correction of operator norm $\rho_\star = 0.10$, $K = 10$ block-constant source covariances), we train two models from the same warm-start A_0 under the same cov-matching loss: LoRA, with $A = A_0 + UV^\top$ at $r = r_\star = 4$, versus a free-A baseline updating all $CN \approx 5 \times 10^4$ entries directly. Both drive the training loss to $\sim 10^{-7}$ (Fig. 2, left), yet the free-A optimum lands at a much larger operator-recovery error $\|A - A_\star\|_F / \|A_\star\|_F$ than LoRA (middle) and at a markedly worse held-out PSNR distribution (right; free A 53.87 dB, A_0 62.37 dB, LoRA 63.89 dB, oracle $A_\star$ 65.22 dB). Without the rank prior, the matching set in unstructured A-space has dimension well above the $K \cdot P(P+1)/2 \approx 1.2 \times 10^4$ available constraints, so the optimiser settles on an arbitrary point in it; the LoRA parameterisation restricts training to a neighbourhood of the target discrete element of $\mathcal{S}_K$. Full setup in Appendix D. Table 1 lists the numbers behind Figure 2. Table 2 in the appendix reports a rank ablation, one run per rank, over $r \in \{1, 2, 4, 8, 16, 32\}$ on this same instance, and Table 5 there reports how the rank of the stacked L_B grows with K.

5 Discussion

Other instances of the template. The setup $(A_0, \Pi, \{\Sigma_y^{(k)}, \Sigma^{\mathrm{ref},(k)}\})$ covers any inverse problem in which two modalities probe a shared hidden source through distinct physical operators with their

Table 1: Readouts for the instance of Figure 2, with both models trained from the same A_0 under the same loss. The recovery errors are reported for the two trained models, and A_0 and the oracle $A_\star$ appear in the PSNR row as reference points.

	A_0	free A	LoRA	oracle $A_\star$
$\|A - A_\star\|_F / \|A_\star\|_F$		0.20	0.018	
$\|B - B_\star\|_F / \|B_\star\|_F$		0.037	0.014	
held-out PSNR (dB)	62.37	53.87	63.89	65.22

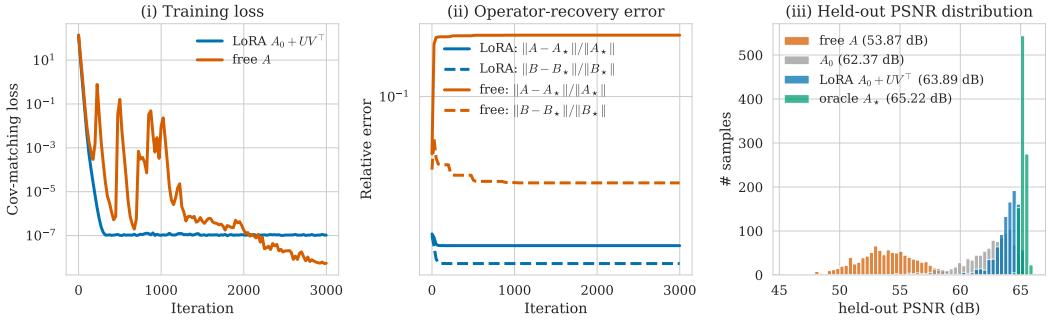


Figure 2: LoRA vs free A under the same cov-matching loss, both warm-started at A_0. (Left) Training loss; both reach $\sim 10^{-7}$. (Middle) Operator-recovery error: solid lines $\|A - A_\star\|_F / \|A_\star\|_F$, dashed lines $\|B - B_\star\|_F / \|B_\star\|_F$ with $B = \Pi A^\dagger$. (Right) Histogram of held-out reconstruction PSNR over 1000 samples; means in legend.

own baselines and aggregations. Joint electromagnetic–seismic geophysical inversion (Gallardo & Meju, 2004; Moorkamp, 2017; Molodtsov et al., 2024; Ma et al., 2025) and gene regulatory network inference (Wang et al., 2024; Moscardó García et al., 2025) are two further instances.

Regime scope. The geometric results of §4.1 (Propositions 3, 4) hold under $\varepsilon_K^* = 0$, i.e. the references are exactly attainable; this is the regime where $\mathcal{S}_K$ carries a Stiefel-manifold continuous degeneracy of dimension $PC - P(P+1)/2$, motivating the multi-reference plus simultaneous-diagonalisation analysis. The gradient-flow results of §4.3 (Theorem 8, Lemma 9, Corollary 10) are stated in terms of the excess loss $\mathcal{L}_K - \varepsilon_K^*$ and of a general bound $D_\|$, so their Hessian argument carries over to $\varepsilon_K^* > 0$, where condition (a) becomes an assumption and the separation δ_B it names is no longer supplied by Proposition 4. Here the residual cross-term $2\sum_k w_k \langle \Gamma_0^{(k)}, \Delta \Sigma_y^{(k)} \Delta^T \rangle$ in (17) is absorbed into condition (b) of Appendix C, and the linear-in-Δ contribution vanishes by stationarity at $B_\star$.

In the $P > C$ single-reference regime, the rank constraint already reduces the solution set to an $O(C)$-orbit of dimension $C(C-1)/2$ (Proposition 1, overdetermined regime), on which the loss is invariant; the loss provides no signal along the orbit, and any orbit point is a valid best fit at the level of second-order moments. The stability results (Lemma 5, Proposition 6, part (1) of Proposition 7) carry over unchanged, so A_0-anchored warm-start retains the trajectory bound $\|B(A_0 + UV^T) - B(A_0)\|_F \le c\rho$ without modification; the orbit point selected by the gradient flow is the one nearest to $B(A_0)$.

Other limitations. The theorems require an available baseline A_0; the case without A_0 is outside the scope of the present paper. The metric on Sym_P^+ is taken as Frobenius; extension to the log-Euclidean metric reduces to Frobenius only in the commuting case $\widehat{\Sigma}\Sigma^{\mathrm{ref}} = \Sigma^{\mathrm{ref}}\widehat{\Sigma}$, while Bures–Wasserstein and affine-invariant metrics require separate Hessian computation and a separate splitting threshold. Proposition 7 assumes the warm-start basin is uniquely determined; the case $\|B(A_0) - B_\star^{(1)}\|_F = \|B(A_0) - B_\star^{(2)}\|_F$ for two distinct $B_\star^{(1)}, B_\star^{(2)} \in \mathcal{S}_K$ is a measure-zero set in the warm-start space.

AI use statement

In this work, we used generative AI tools (Claude Opus, Anthropic) for the following tasks with required disclosure. We used them to generate the synthetic data sets of the numerical experiments. We used them to read and digest the literature, to adapt the key mathematical tools to our setting, and to assist in deriving and verifying the theoretical results, including the writing of proofs. We used them to assist in designing the experiments, implementing the methods, analysing and interpreting the results, and drafting the experiment reports. We have not used generative AI tools for translation. Cleaning or reformatting data sets and qualitative or thematic data analysis are not applicable to this work, since all data are synthetic. Additionally, we used generative AI tools for the pen-and-ink stylisation of Figure 1, whose geometric layout was computed programmatically. We have reviewed all AI-assisted work. Every proof produced with AI assistance was checked and verified by hand on paper, and the code and data behind the figures were checked and cross-validated. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This work is theoretical. Its experiments use synthetic coupling matrices, and involve no human subjects or personal data. We are not aware of any ethical concern raised by this work.

Reproducibility statement

Every hypothesis used by the results is listed, with the results that use it, in Appendix E. Complete proofs are in Appendices A, B and C. The data generation, the training configuration and the full numerical results are in Appendix D, and the code is available in the supplementary material.

References

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Beyond Pham's algorithm for joint diagonalization, 2019.

Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017. doi: 10.1088/1361-6420/aa9581.

Abdeldjallil Aissa El Bey, Karim Abed-Meraim, Yves Grenier, and Yingbo Hua. A general framework for second order blind separation of stationary colored sources. *Signal Processing*, 88(9): 2123–2137, 2008. doi: 10.1016/j.sigpro.2008.03.017.

Simon Arridge, Andreas Hauptmann, and Yury Korolev. Inverse problems with learned forward operators, 2024. v2, March 2024.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997. doi: 10.1109/78.554307.

Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996. doi: 10.1137/S0895479893259546.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H. Salat, Evelina Busa, Larry J. Seidman, Jill Goldstein, David Kennedy, Verne Caviness, Nikos Makris, Bruce Rosen, and Anders M. Dale. Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22, 2004. doi: 10.1093/cercor/bhg087.

Luis A. Gallardo and Max A. Meju. Joint two-dimensional DC resistivity and seismic travel time inversion with cross-gradients constraints. *Journal of Geophysical Research: Solid Earth*, 109 (B3):B03311, 2004. doi: 10.1029/2003JB002716.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973. doi: 10.1137/0710036.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Nicholas Krämer, Pablo Moreno-Muñoz, Hrittik Roy, and Søren Hauberg. Gradients of functions of large matrices. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Spotlight.

Jian Ma, Yuxiao Deng, Xinming Li, et al. Recent advances in machine learning-enhanced joint inversion of seismic and electromagnetic data. *Surveys in Geophysics*, 46:197–225, 2025. doi: 10.1007/s10712-024-09867-3.

Dmitry Molodtsov, Duygu Kiyan, and Christopher J. Bean. Decoupled joint inversion with variable splitting: example scheme for magnetotelluric, seismic and gravity data. *Geophysical Journal International*, 239(1):706–724, 2024. doi: 10.1093/gji/ggae292.

Max Moorkamp. Integrating electromagnetic data with other geophysical observations for enhanced imaging of the earth: A tutorial and review. *Surveys in Geophysics*, 38(5):935–962, 2017. doi: 10.1007/s10712-017-9413-7.

María Moscardó García, Atte Aalto, Arthur N. Montanari, Alexander Skupin, and Jorge Gonçalves. Multi-omic network inference from time-series data. *npj Systems Biology and Applications*, 11 (1):114, 2025. doi: 10.1038/s41540-025-00591-1.

John C. Mosher, Richard M. Leahy, and Paul S. Lewis. EEG and MEG: forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259, 1999. doi: 10.1109/10.748978.

Rossen Nenov, Daniel Haider, and Peter Balazs. (almost) smooth sailing: Towards numerical stability of neural networks through differentiable regularization of the condition number. In *ICML Workshop on Differentiable Almost Everything*, 2024.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, Rebecca Willett, and Mahdi Soltanolkotabi. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

Yan Pan, Markus Matilainen, Sara Taskinen, and Klaus Nordhausen. A review of second-order blind identification methods. *WIREs Computational Statistics*, 2022. doi: 10.1002/wics.1550.

Théodore Papadopoulo and Manolis Lourakis. Estimating the Jacobian of the singular value decomposition: Theory and applications. In *European Conference on Computer Vision (ECCV)*, volume 1842 of *Lecture Notes in Computer Science*, pp. 554–570. Springer, 2000. doi: 10.1007/3-540-45054-8_36.

Dinh Tuan Pham. Joint approximate diagonalization of positive definite Hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1136–1152, 2001. doi: 10.1137/S089547980035689X.

Álvaro Ribot, Anna Seigal, and Piotr Zwiernik. Beyond independent component analysis: Identifiability and algorithms. *arXiv preprint*, arXiv:2510.07525, 2025.

Yue Song, Nicu Sebe, and Wei Wang. Fast differentiable matrix square root and inverse square root. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. doi: 10.1109/TPAMI.2022.3216339. Extended journal version of the ICLR 2022 paper "Fast Differentiable Matrix Square Root" (arXiv:2201.08663).

G. W. Stewart. On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, 17(1):33–45, 1969. doi: 10.1137/0117004.

Petr Tichavský and Arie Yeredor. Fast approximate joint diagonalization incorporating weight matrices. *IEEE Transactions on Signal Processing*, 57(3):878–891, 2009. doi: 10.1109/TSP.2008.2009271.

Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2520–2523, 2011. doi: 10.1109/ICASSP.2011.5946997.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, and Chenfeng Xu. Dobi-SVD: Differentiable SVD for LLM compression and some new perspectives. In *International Conference on Learning Representations (ICLR)*, 2025.

Wei Wang, Zheng Dang, Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Robust differentiable SVD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5472–5487, 2022. doi: 10.1109/TPAMI.2021.3072422.

Yue Wang, Peng Zheng, Yu-Chen Cheng, Zikun Wang, and Aleksandr Aravkin. WENDY: Covariance dynamics based gene regulatory network inference. *Mathematical Biosciences*, 377:109284, 2024. doi: 10.1016/j.mbs.2024.109284.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. doi: 10.1007/BF01932678.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973. doi: 10.1007/BF01933494.

Yinghao Zhang and Yue Hu. Differentiable SVD based on Moore–Penrose pseudoinverse for inverse imaging problems, 2024.

A Proofs for §4.1: the matching set in B-space

Proof of Proposition 2

The rank condition implies the counting condition. The codomain of $L_{B_\star}$ has dimension $\sum_k n_k$, so rank $L_{B_\star} \leq \sum_k n_k$, and rank $L_{B_\star} = PC$ forces $\sum_k n_k \geq PC$.

The rank condition is sufficient for discreteness. Under $\varepsilon_K^* = 0$ the matching set is the zero set of $F : B \mapsto \left(B\Sigma_y^{(k)}B^T - \Sigma^{\mathrm{ref},(k)}\right)_{k=1,\dots,K}$, whose differential at $B_\star$ is $L_{B_\star}$. The rank condition makes that differential injective, so F is an immersion at $B_\star$ and is injective on a neighbourhood of $B_\star$. Since $F(B_\star) = 0$, no other point of that neighbourhood belongs to $\mathcal{S}_K$, and $B_\star$ is isolated in $\mathcal{S}_K$.

The converse implication, from discreteness at $B_\star$ to the counting condition, is false. The example at the end of §A.2 has $\sum_k n_k < PC$ and a matching set of two isolated points.

A.1 Proof of Proposition 3

Throughout this subsection $m := \min(P, C)$ and $\Psi := \Lambda^{-1/2}Q^T$, as in §4.1 of the main text, and

$$G_k \;:=\; \Sigma_y^{-1/2}\,\Sigma_y^{(k)}\,\Sigma_y^{-1/2} \;\in\; \mathbb{R}^{C\times C}, \qquad k = 1, \dots, K, \tag{20}$$

where $\Sigma_y^{1/2}$ is the symmetric positive-definite square root of Σ_y. The base pair $(\Sigma_y, \Sigma^{\mathrm{ref}})$ is one of the K pairs, say the one with index k_0, so that $G_{k_0} = I_C$, $D_{k_0} = I_P$ and $d_{ik_0} = 1$ for every i; the base equation is the equation of index k_0, and we write it separately where that is convenient. We assume that the matching equations have at least one solution; otherwise the statement holds vacuously.

Step 1: whitening. Assume first $P \le C$, so that $m = P$ and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_P)$ is invertible. The map $B \mapsto \widetilde{B} := \Psi B$ is a linear bijection of $\mathbb{R}^{P\times C}$ with inverse $\widetilde{B} \mapsto Q\Lambda^{1/2}\widetilde{B}$. Substituting $B = Q\Lambda^{1/2}\widetilde{B}$ and using $Q^TQ = QQ^T = I_P$,

$$B\Sigma_y B^T = \Sigma^{\mathrm{ref}} \iff Q\Lambda^{1/2}\,\widetilde{B}\Sigma_y\widetilde{B}^T\,\Lambda^{1/2}Q^T = Q\Lambda Q^T \iff \widetilde{B}\Sigma_y\widetilde{B}^T = I_P,$$
$$B\Sigma_y^{(k)} B^T = \Sigma^{\mathrm{ref},(k)} \iff \widetilde{B}\Sigma_y^{(k)}\widetilde{B}^T = \Lambda^{-1/2}Q^T\Sigma^{\mathrm{ref},(k)}Q\Lambda^{-1/2} = D_k.$$

Put $M := \widetilde{B}\Sigma_y^{1/2}$, again a linear bijection of $\mathbb{R}^{P\times C}$. The two families of equations become

$$MM^T = \widetilde{B}\Sigma_y\widetilde{B}^T = I_P, \qquad MG_kM^T = \widetilde{B}\Sigma_y^{(k)}\widetilde{B}^T = D_k. \tag{21}$$

The task is to find every M with orthonormal rows that satisfies $MG_kM^T = D_k$ for $k = 1, \dots, K$.

Step 2: the square case $P = C$. Here M is square with $MM^T = I_C$, so $M^T = M^{-1}$. Multiplying $MG_kM^T = D_k$ on the right by M gives $MG_k = D_kM$. Reading row i of both sides, with m_i^T the i-th row of M,

$$m_i^T G_k = d_{ik}\, m_i^T, \qquad i = 1, \dots, C, \quad k = 1, \dots, K, \tag{22}$$

so m_i is a common left eigenvector of $G_1, \dots, G_K$ whose eigenvalue profile is $\boldsymbol{\rho}_i = (d_{i1}, \dots, d_{iK})$. Define

$$\mathcal{V}_i := \{v \in \mathbb{R}^C : v^T G_k = d_{ik}\, v^T \text{ for } k = 1, \dots, K\}.$$

By hypothesis the profiles are pairwise distinct.

We claim that the sum $\mathcal{V}_1 + \cdots + \mathcal{V}_C$ is direct. Suppose it is not, and choose a relation $\sum_{i\in S} v_i = 0$ with $v_i \in \mathcal{V}_i \setminus \{0\}$ and S of minimal cardinality among all such relations, so that $|S| \ge 2$. Fix $i \ne j$ in S and an index k with $d_{ik} \ne d_{jk}$. Applying G_k^T to the relation gives $\sum_{i'\in S} d_{i'k}\, v_{i'} = 0$, and subtracting d_{jk} times the relation itself gives

$$\sum_{i'\in S} (d_{i'k} - d_{jk})\, v_{i'} = 0.$$

The term $i' = j$ has coefficient zero and the term $i' = i$ has coefficient different from zero, so this is a relation of the same kind whose index set is non-empty and strictly smaller than S, which contradicts minimality.

The C rows of M are linearly independent and $m_i \in \mathcal{V}_i$, so $\dim \mathcal{V}_i \ge 1$ for every i, while, the sum being direct, $\sum_i \dim \mathcal{V}_i \le C$. Hence $\dim \mathcal{V}_i = 1$ for every i. Let M and M' both solve (21). Their i-th rows are unit vectors of the one-dimensional space $\mathcal{V}_i$, so $m_i' = \varepsilon_i m_i$ with $\varepsilon_i \in \{\pm 1\}$, that is $M' = D_\pm M$ with $D_\pm = \mathrm{diag}(\varepsilon_1, \dots, \varepsilon_C)$. Conversely each $D_\pm M$ solves (21), since $(D_\pm M)(D_\pm M)^T = D_\pm MM^T D_\pm = I_C$ and $(D_\pm M)G_k(D_\pm M)^T = D_\pm D_k D_\pm = D_k$ by commutation of diagonal matrices. The solution set has $2^C = 2^m$ elements. Undoing the two substitutions, $B = Q\Lambda^{1/2}M\Sigma_y^{-1/2}$, and $M' = D_\pm M$ corresponds to

$$B' = Q\Lambda^{1/2}D_\pm M\Sigma_y^{-1/2} = \left(Q\Lambda^{1/2}D_\pm\Lambda^{-1/2}Q^T\right)B, \tag{23}$$

which is the action in the statement.

Step 3: the overdetermined case $P > C$. Here $m = C$ and Λ is singular, so Ψ is defined on the range of Σ^{ref}, as follows. For any solution B we have $\Sigma^{\mathrm{ref}} = (B\Sigma_y^{1/2})(B\Sigma_y^{1/2})^T$, so $\mathrm{rank}\,\Sigma^{\mathrm{ref}} = \mathrm{rank}\,B \le C$, while the hypothesis $\lambda_C > 0$ gives $\mathrm{rank}\,\Sigma^{\mathrm{ref}} \ge C$. Therefore $\mathrm{rank}\,\Sigma^{\mathrm{ref}} = \mathrm{rank}\,B = C$ and $\lambda_{C+1} = \cdots = \lambda_P = 0$. In this regime the matching equations are solvable only for a singular Σ^{ref} of rank exactly C, which agrees with the overdetermined branch of Proposition 1; the convention $\Sigma^{\mathrm{ref}} \in \mathrm{Sym}_P^+$ of the notation table is relaxed here to positive semi-definiteness of rank C.

Write the rank decomposition $\Sigma^{\mathrm{ref}} = Q_C\Lambda_C Q_C^T$ with $Q_C \in \mathbb{R}^{P\times C}$ of orthonormal columns and $\Lambda_C = \mathrm{diag}(\lambda_1, \dots, \lambda_C)$ positive definite, and set $\Psi := \Lambda_C^{-1/2}Q_C^T$ and $D_k := \Psi\Sigma^{\mathrm{ref},(k)}\Psi^T$, both of size $C \times C$; for $P \le C$ these are the definitions of the main text. The identity $\Sigma^{\mathrm{ref}} = (B\Sigma_y^{1/2})(B\Sigma_y^{1/2})^T$ also gives $\mathrm{col}\,\Sigma^{\mathrm{ref}} = \mathrm{col}\,B$, so $Q_CQ_C^TB = B$ and $B = Q_CX$ with

$X := Q_C^T B \in \mathbb{R}^{C\times C}$ invertible. Substituting in the base equation and multiplying by Q_C^T on the left and by Q_C on the right,

$$Q_C X \Sigma_y X^T Q_C^T \;=\; Q_C \Lambda_C Q_C^T \iff X \Sigma_y X^T \;=\; \Lambda_C,$$

and in the same way $X\Sigma_y^{(k)}X^T = Q_C^T \Sigma^{\mathrm{ref},(k)} Q_C$ for each k; the second family is an equivalence because $\Sigma^{\mathrm{ref},(k)} = B\Sigma_y^{(k)}B^T$ has its column space and its row space inside $\operatorname{col} Q_C$. With $\widetilde{B} := \Psi B = \Lambda_C^{-1/2} X$ this reads $\widetilde{B}\Sigma_y\widetilde{B}^T = I_C$ and $\widetilde{B}\Sigma_y^{(k)}\widetilde{B}^T = D_k$, the square system of Step 2 in dimension C. Its solution set has $2^C = 2^m$ elements related by $\widetilde{B} \mapsto D_\pm \widetilde{B}$, and undoing the substitution gives the action $B \mapsto Q_C \Lambda_C^{1/2} D_\pm \Lambda_C^{-1/2} Q_C^T B$, using $Q_C^T Q_C = I_C$. Steps 2 and 3 use neither (6) nor (7).

Step 4: the underdetermined case $P < C$. The argument of Step 2 is unavailable here, because $MM^T = I_P$ does not give $M^TM = I_C$ and therefore $MG_kM^T = D_k$ does not give $MG_k = D_kM$. The matching equations constrain the compression of each G_k to the row space of M and leave that row space itself free; the remark at the end of this subsection exhibits a continuum of solutions under the hypotheses of the proposition alone. We use the two further hypotheses stated in the main text at (6) and (7). By (6) there are $\Omega \in O(C)$ and scalars γ_{ck} with $G_k = \Omega\,\Xi_k\,\Omega^T$ and $\Xi_k := \operatorname{diag}(\gamma_{1k},\dots,\gamma_{Ck})$, and the channel profile $\gamma_c := (\gamma_{c1},\dots,\gamma_{cK}) \in \mathbb{R}^K$ collects the c-th diagonal entries. By (7) each $\boldsymbol{\rho}_i$ with $i \le P$ is a vertex, that is an extreme point, of $\operatorname{conv}\{\gamma_1,\dots,\gamma_C\}$ and equals γ_c for exactly one index c, written $c(i)$.

Let M solve (21) and set $X := M\Omega \in \mathbb{R}^{P\times C}$, whose i-th row we write as x_i^T. Then $XX^T = M\Omega\Omega^TM^T = I_P$ and $X\Xi_kX^T = M\Omega\Xi_k\Omega^TM^T = MG_kM^T = D_k$. Reading the i-th diagonal entry of both identities,

$$\sum_{c=1}^{C} x_{ic}^2 \;=\; 1, \qquad \sum_{c=1}^{C} \gamma_{ck}\, x_{ic}^2 \;=\; d_{ik}, \qquad k = 1,\dots,K. \tag{24}$$

Set $p_{ic} := x_{ic}^2 \ge 0$. The first identity says that $p_i := (p_{i1},\dots,p_{iC})$ lies in the simplex of probability vectors on C points, and the second says $\sum_c p_{ic}\,\gamma_c = \boldsymbol{\rho}_i$. In particular $\boldsymbol{\rho}_i$ lies in $\operatorname{conv}\{\gamma_1,\dots,\gamma_C\}$ for every solution, so (7) adds only extremality and the uniqueness of the index.

Let $T_i := \{c : p_{ic} > 0\}$, which is non-empty. If $T_i = \{c\}$ then $\boldsymbol{\rho}_i = \gamma_c$ and $c = c(i)$ by the uniqueness in (7). Suppose $|T_i| \ge 2$ and take $c \in T_i$. With $q := p_{ic} \in (0,1)$ and

$$z \;:=\; \frac{1}{1-q}\sum_{c'\neq c} p_{ic'}\,\gamma_{c'} \;\in\; \operatorname{conv}\{\gamma_1,\dots,\gamma_C\},$$

one has $\boldsymbol{\rho}_i = q\,\gamma_c + (1-q)\,z$. An extreme point of a convex set is not an interior point of a segment joining two points of that set, so $\gamma_c = z = \boldsymbol{\rho}_i$. This holds for every $c \in T_i$ and therefore produces at least two indices c with $\gamma_c = \boldsymbol{\rho}_i$, which contradicts the uniqueness in (7). Hence $T_i = \{c(i)\}$, $p_{i,c(i)} = 1$, and $x_i = \varepsilon_i e_{c(i)}$ with $\varepsilon_i \in \{\pm 1\}$.

The indices $c(1),\dots,c(P)$ are distinct, since $\gamma_{c(i)} = \boldsymbol{\rho}_i$ and the profiles $\boldsymbol{\rho}_i$ are pairwise distinct. Every matrix of this form is a solution: writing $E := [\,e_{c(1)}\ \cdots\ e_{c(P)}\,]^T$ and $X = D_\pm E$ with $D_\pm \in \{\pm1\}^P$ diagonal, the rows of X have disjoint supports, so $XX^T = I_P$, and $(X\Xi_kX^T)_{ij} = \varepsilon_i\varepsilon_j\,\gamma_{c(i),k}\,\delta_{ij}$, which is D_k because $\gamma_{c(i),k} = d_{ik}$. The solution set in the variable X therefore consists of the $2^P = 2^m$ matrices $D_\pm E$. Undoing the substitutions, $M = X\Omega^T$ and $B = Q\Lambda^{1/2}X\Omega^T\Sigma_y^{-1/2}$, and two solutions $X = D_\pm E$ and $X' = D'_\pm E$ satisfy $X' = (D'_\pm D_\pm)X$, so the corresponding operators satisfy $B' = Q\Lambda^{1/2}(D'_\pm D_\pm)\Lambda^{-1/2}Q^TB$, the action in the statement.

Step 5: the prescribed diagonal and the permutation factor. In the joint blind source separation setting cited in §2 of the main text, the diagonal values are not prescribed, the solution set is closed under permutation of the rows, and the indeterminacy has order $2^m m!$. Here the right-hand sides D_k are supplied by the data, so row i of $\widetilde{B}$ has to reproduce the prescribed profile $\boldsymbol{\rho}_i$. Composing a solution with a row permutation $\pi \neq \mathrm{id}$ gives row i the profile $\boldsymbol{\rho}_{\pi(i)}$, which differs from $\boldsymbol{\rho}_i$, so the (i,i) entry of the k-th equation fails for some k. The permutation factor is removed in this way and

only the 2^m sign choices remain. The strict ordering $\lambda_1 > \cdots > \lambda_m > 0$ does the same for the base equation: it fixes the columns of Q up to sign, hence fixes the row labels in the un-whitened coordinates.

Remark: hypothesis (7) cannot be dropped. Take $P = 1$, $C = 2$ and $K = 2$ pairs, namely the base pair $(\Sigma_y, \Sigma^{\text{ref}}) = (I_2, 1)$ and $(\Sigma_y^{(2)}, \Sigma^{\text{ref},(2)}) = (aI_2, a)$ with $a > 0$. Every hypothesis of Proposition 3 holds: $m = 1$, $\lambda_1 = 1 > 0$, both D_k are diagonal, and pairwise distinctness is vacuous for one profile. Hypothesis (6) holds as well, with $\Omega = I_2$ and the two diagonal factors I_2 and aI_2. Writing $B = (x, y)$, the matching equations are $x^2 + y^2 = 1$ and $a(x^2 + y^2) = a$, and the second is implied by the first, so the solution set is the whole unit circle instead of $2^m = 2$ points. What fails is (7): $\gamma_1 = \gamma_2 = (1, a)$, so the index $c(1)$ is not unique.

Remarks on Proposition 3. Proposition 3 holds for every $B \in \mathbb{R}^{P\times C}$ that satisfies the K matching equations; no neighbourhood of a solution, no A_0 and no warm start enters its hypotheses. What the data leave undetermined is the group of row-wise sign flips in the whitened coordinates, a subgroup of the group of sign, permutation and scale indeterminacies that is standard in second-order source separation. All $2^{\min(P,C)}$ elements of the orbit produce the same $\widehat{\Sigma}^{(k)}$ for every k, so second-order data of this kind cannot separate them. The count $2^{\min(P,C)}$ does not depend on K; what K decides is whether the hypotheses can hold at all. Below the counting threshold of Proposition 2 the linearised system is deficient, and above it further references improve conditioning without changing the solution set.

The vector $\boldsymbol{\rho}_i$ collects the i-th diagonal entry of D_k across $k = 1, \ldots, K$, so it records how the i-th whitened row responds over the K conditions. Distinctness of the profiles is what makes the residual group finite: with $M := \widetilde{B}\,\Sigma_y^{1/2}$ the matching equations read $MM^T = I_P$ and $MG_kM^T = D_k$, and ΘM solves the same pair for $\Theta \in O(P)$ exactly when Θ commutes with every D_k. That commutant consists of the sign matrices $D_\pm \in \{\pm 1\}^{\min(P,C)}$ when the profiles are pairwise distinct, and contains a full $O(2)$ block, hence a continuous curve of loss-equivalent B, as soon as two profiles coincide. Pairwise non-proportionality is the form the hypothesis takes in Aissa El Bey et al. (2008), where the diagonal is not prescribed; here the base pair contributes the entry 1 to every profile, so the two forms agree.

Condition (6) asks the whitened channel covariances to be diagonalised by one common orthogonal matrix, which for symmetric matrices is equivalent to their pairwise commutation. Membership of $\boldsymbol{\rho}_i$ in the hull is automatic for any B that solves the matching equations, so (7) asks only for the vertex property, and the remark on (7) above shows that it cannot be dropped. For $P \geq C$ neither hypothesis is needed, since Steps 2 and 3 above complete the reduction without them. Both conditions, and the diagonality of D_k, require the references to be symmetric; the non-symmetric lag references enter only the counting condition of Proposition 2.

A.2 The kernel of $L_{B_\star}$ and the counting condition

Throughout this subsection $P < C$, the hypotheses of Proposition 3 hold together with (6) and (7), and $\varepsilon_K^* = 0$. As in §A.1, the base pair carries the index k_0, so that $\gamma_{c,k_0} = 1$ for every c and the k_0 entry of every profile is 1.

In the coordinates of Step 4 of §A.1 the map $B \mapsto B\Sigma_y^{(k)}B^T$ reads $X \mapsto \Psi^{-1}(X\Xi_kX^T)\Psi^{-T}$, where $B = \Psi^{-1}X\Omega^T\Sigma_y^{-1/2}$ and Ψ, Ω, $\Sigma_y^{1/2}$ are fixed invertible matrices. Both $B \mapsto X$ and $Z \mapsto \Psi^{-1}Z\Psi^{-T}$ are linear bijections, so $\ker L_{B_\star}$ is mapped isomorphically onto

$$\mathcal{K} \;:=\; \big\{\xi \in \mathbb{R}^{P\times C} : \xi\Xi_kX^T + X\Xi_k\xi^T = 0,\; k = 1, \ldots, K\big\},$$

taken at the solution X that corresponds to $B_\star$. We may take $X = E$, because replacing X by $D_\pm X$ replaces ξ by $D_\pm\xi$ and leaves the dimension unchanged. The i-th row of E is $e_{c(i)}^T$, so

$$(\xi\Xi_kE^T)_{ij} \;=\; \gamma_{c(j),k}\,\xi_{i,c(j)}, \qquad (E\Xi_k\xi^T)_{ij} \;=\; \gamma_{c(i),k}\,\xi_{j,c(i)},$$

and $\mathcal{K}$ is the set of ξ satisfying

$$\gamma_{c(j),k}\,\xi_{i,c(j)} \;+\; \gamma_{c(i),k}\,\xi_{j,c(i)} \;=\; 0 \tag{25}$$

for all $i, j \in \{1, \dots, P\}$ and all $k \in \{1, \dots, K\}$. At $i = j$ and $k = k_0$ this reads $2\,\xi_{i,c(i)} = 0$. For $i \neq j$ the pair $(\xi_{i,c(j)}, \xi_{j,c(i)})$ solves a homogeneous system in two unknowns whose coefficient columns are $(\gamma_{c(j),k})_{k=1}^{K} = \boldsymbol{\rho}_j$ and $(\gamma_{c(i),k})_{k=1}^{K} = \boldsymbol{\rho}_i$. Suppose these two vectors were proportional. Their k_0 entries are both 1, so the factor would be 1, and then $\boldsymbol{\rho}_i = \boldsymbol{\rho}_j$, which is excluded. They are therefore not proportional, the system has rank two, and $\xi_{i,c(j)} = \xi_{j,c(i)} = 0$. Every entry ξ_{ic} with $c \in \{c(1), \dots, c(P)\}$ vanishes, and (25) places no condition on the entries with c outside that set. Counting the free entries,

$$\dim \ker L_{B_\star} \;=\; P\,(C - P). \tag{26}$$

Two consequences follow. First, rank $L_{B_\star} = PC - P(C-P) = P^2 < PC$, so under the hypotheses of this subsection the rank condition of Proposition 2 does not hold, while Proposition 3 gives a matching set of 2^P isolated points. In the underdetermined regime with exact matching, discreteness is established by the extreme-point argument of Step 4 above, and not by the rank of $L_{B_\star}$.

Second, the counting condition $\sum_k n_k \geq PC$ is not necessary for discreteness. Take $P = 1$, $C = 3$ and $K = 2$ symmetric pairs, one of which is the base pair, so that $n_1 = n_2 = 1$ and $\sum_k n_k = 2 < 3 = PC$. Choose the data so that, in the notation above, the two diagonal factors are I_3 and $\mathrm{diag}(a_1, a_2, a_3)$ with a_1, a_2, a_3 pairwise distinct and $a_1 < \min(a_2, a_3)$, and let the second pair prescribe $\boldsymbol{\rho}_1 = (1, a_1)$. Writing $X = (x, y, z)$, the matching equations are $x^2 + y^2 + z^2 = 1$ and $a_1 x^2 + a_2 y^2 + a_3 z^2 = a_1$; subtracting a_1 times the first from the second gives $(a_2 - a_1)y^2 + (a_3 - a_1)z^2 = 0$, hence $y = z = 0$ and $x = \pm 1$. The matching set is the pair of isolated points $\{\pm e_1\}$ although the counting condition fails. Only the implication from the rank condition to the counting condition holds, as proved above for Proposition 2.

Proof of Proposition 4

Let B_0 and B_1 be two distinct elements of $\mathcal{S}_K$. By Proposition 3 their whitened forms $\widetilde{B}_0 = \Psi B_0$ and $\widetilde{B}_1 = \Psi B_1$ satisfy $\widetilde{B}_1 = D_\pm \widetilde{B}_0$ with $D_\pm = \mathrm{diag}(\varepsilon_1, \dots, \varepsilon_m)$ and $D_\pm \neq I_m$; let F be the non-empty set of indices i with $\varepsilon_i = -1$. Write $\widetilde{b}_i^T$ for the i-th row of $\widetilde{B}_0$. The i-th diagonal entry of $\widetilde{B}_0 \Sigma_y \widetilde{B}_0^T = I_m$ gives $\widetilde{b}_i^T \Sigma_y \widetilde{b}_i = 1$, and $\widetilde{b}_i^T \Sigma_y \widetilde{b}_i \leq \sigma_{\max}(\Sigma_y)\, \|\widetilde{b}_i\|_2^2$, so $\|\widetilde{b}_i\|_2^2 \geq 1/\sigma_{\max}(\Sigma_y)$. The matrix $\widetilde{B}_0 - \widetilde{B}_1$ has i-th row $2\,\widetilde{b}_i^T$ for $i \in F$ and zero rows otherwise. Un-whitening multiplies row i by $\lambda_i^{1/2}$ and then applies Q, which has orthonormal columns and hence preserves the Frobenius norm, so

$$\|B_0 - B_1\|_F^2 \;=\; \sum_{j \in F} 4\,\lambda_j\, \|\widetilde{b}_j\|_2^2 \;\geq\; \frac{4\,\lambda_i}{\sigma_{\max}(\Sigma_y)}$$

for each $i \in F$, which is (8). Since F is non-empty and $\lambda_i \geq \lambda_m$ for every $i \leq m$, the bound $\|B_0 - B_1\|_F \geq \delta_B$ follows. No choice of representatives is involved, because every sign class of $\mathcal{S}_K$ is a single point under the hypotheses of Proposition 3.

A.3 Exact attainability of the references

Let $A \in \mathbb{R}^{C \times N}$ have full row rank, so that $A^\dagger A = A^T (A A^T)^{-1} A$ is the orthogonal projector of $\mathbb{R}^N$ onto $\mathrm{row}(A)$, and suppose $\mathrm{row}(\Pi) \subseteq \mathrm{row}(A)$. Each row of Π is then fixed by that projector, and the projector is symmetric, so

$$\Pi A^\dagger A \;=\; \Pi. \tag{27}$$

Let $\Sigma_x^{(k)}$ be the source covariance on segment k, and let the two observations be $y = Ax$ and Πx, so that $\Sigma_y^{(k)} = A \Sigma_x^{(k)} A^T$ and $\Sigma^{\mathrm{ref},(k)} = \Pi \Sigma_x^{(k)} \Pi^T$. With $B := \Pi A^\dagger$ and (27),

$$B \Sigma_y^{(k)} B^T \;=\; (\Pi A^\dagger A)\, \Sigma_x^{(k)}\, (\Pi A^\dagger A)^T \;=\; \Pi \Sigma_x^{(k)} \Pi^T \;=\; \Sigma^{\mathrm{ref},(k)}$$

for every k and every $\Sigma_x^{(k)}$, so $\varepsilon_K^* = 0$ for every K and every choice of segments. Exact attainability is in this way a property of the data and of the row space of A.

The instance of §D.1 is built this way. There A_0 is formed by stacking the row-normalised Π above $C - P$ further rows, so $\mathrm{row}(\Pi) \subseteq \mathrm{row}(A_0)$, and $A_\star = (I_C + J) A_0$ with $I_C + J$ invertible, so $\mathrm{row}(A_\star) = \mathrm{row}(A_0)$ and (27) holds at $A_\star$. This is why the $K = 10$ references of that instance are simultaneously attainable.

B Proofs for §4.2: the pseudo-inverse differential and rank dependence

Proof of Lemma 5

Bound each term of (10) using $\|XYZ\|_F \le \|X\|_{\text{op}}\|Y\|_F\|Z\|_{\text{op}}$. For the first term, $\|A^\dagger\|_{\text{op}} = \sigma_{\min}(A)^{-1}$ gives $\|A^\dagger(\Delta A)A^\dagger\|_F \le \sigma_{\min}(A)^{-2}\,\|\Delta A\|_F$. For the second, $\|I_N - A^\dagger A\|_{\text{op}} = 1$ and $\|(AA^T)^{-1}\|_{\text{op}} = \sigma_{\min}(A)^{-2}$ give the same bound. Adding the two contributions yields (9).

Proof of Proposition 6

$UV^T = \sum_i u_i v_i^T$, so by the triangle and sub-multiplicative inequalities $\|UV^T\|_{\text{op}} \le \sum_i \|u_i\|_2\|v_i\|_2 = r\rho^2$. Substituting $\sigma_{\min}(A) \ge \sigma_{\min}(A_0) - r\rho^2$ into Lemma 5 produces $\bar\kappa(r)$.

Proof of Proposition 7

(1) Apply Lemma 5 along the segment $A(s) := A_0 + s\,UV^T$, $s \in [0,1]$. Weyl gives $\sigma_{\min}(A(s)) \ge \sigma_0 - \rho \ge \sigma_0/2$, hence $2/\sigma_{\min}(A(s))^2 \le 8/\sigma_0^2$. Integrating and using $\|UV^T\|_F \le \sqrt{r}\,\|UV^T\|_{\text{op}} \le \sqrt{r}\,\rho$ for $\operatorname{rank}(UV^T) \le r$, then composing with Π, gives $c\rho$ with the stated c. (2) Triangle inequality with (14) and (15) places $B(A_0 + UV^T)$ within $\delta_B/2$ of $B_\star$; Proposition 4 separates other sign-classes by δ_B. (3) Frobenius-projection non-expansiveness on $\|B_t - B_\star\|_F \le c\rho + \|B(A_0) - B_\star\|_F$.

C Proofs for §4.3: gradient flow near A_0

Conditions for Corollary 10

The following conditions are assumed throughout §4.3 of the main text.

(a) The simultaneous-diagonalisation hypothesis of Proposition 3 holds: all reference covariances $\Sigma^{\text{ref},(k)}$ share the same eigenbasis, together with (6) and (7) of the main text when $P < C$, so $\mathcal{S}_K$ is a discrete set of sign-class representatives separated in Frobenius norm by δ_B (Proposition 4).

(b) At $B_\star \in \mathcal{S}_K$, the second-order term H of (17) is positive-definite on $\ker L_{B_\star}^{\perp}$. Write

$$h \;:=\; \min\big\{H[\Delta] : \Delta \in \ker L_{B_\star}^{\perp},\ \|\Delta\|_F = 1\big\}, \qquad S \;:=\; \sum_k w_k \|\Sigma_y^{(k)}\|_{\text{op}}^2,$$

so that $H[\Delta] \ge h\,\|\Delta\|_F^2$ on $\ker L_{B_\star}^{\perp}$. The quantitative requirement on h is

$$h \;\ge\; 12\|B_\star\|_{\text{op}}\, S\,\delta_B \;+\; \tfrac{3S\,\delta_B^2}{2}. \tag{28}$$

This guarantees that the trust-region radius r_0 in the Taylor analysis of Lemma 9 satisfies $r_0 \ge \delta_B/2$; the derivation is at the end of the proof of Lemma 9 below.

(c) The gradient-flow trajectory stays in $\mathcal{F}_A$ and $B_{t_0} \in \mathcal{B}(B_\star, \delta_B/2)$ and $\mathcal{L}_K(B_{t_0}) - \varepsilon_K^* < \alpha^\star$ at some finite $t_0 \ge 0$, where $\alpha^\star$ is the excess-loss threshold (18) of Lemma 9. The two requirements are checked differently. Under the warm-start hypothesis (15), Proposition 7 (2) gives the first one at every admissible (U, V), and with a strict inequality, since $\|B(A_0 + UV^T) - B_\star\|_F \le c\rho + \|B(A_0) - B_\star\|_F < \delta_B/2$. The second requirement concerns the value of the loss and is checked during training. The hypothesis (15) bounds a distance in B-space and gives no upper bound on $\mathcal{L}_K$, so it does not imply the second requirement.

Conditions (a)–(b) depend only on $(A_0, \Pi, \{\Sigma_y^{(k)}\}, \{\Sigma^{\text{ref},(k)}\}, B_\star)$ and are verifiable before training. Of the two requirements in condition (c), the containment holds a priori under the warm start, while the bound on the loss has to be checked during training.

Proof of Theorem 8

Gradient. With $\Gamma^{(k)} := \widehat{\Sigma}^{(k)}(B) - \Sigma^{\mathrm{ref},(k)}$ and $\mathcal{L}^{(k)} := \|\Gamma^{(k)}\|_F^2$, so that $\mathcal{L}_K = \sum_k w_k \mathcal{L}^{(k)}$ as in (5), the chain rule gives

$$\frac{\partial \mathcal{L}^{(k)}}{\partial B} \;=\; 2(\Gamma^{(k)})^T B \Sigma_y^{(k)} \;+\; 2\Gamma^{(k)} B (\Sigma_y^{(k)})^T, \tag{29}$$

which reduces to $4\Gamma^{(k)} B \Sigma_y^{(k)}$ for symmetric $\Sigma_y^{(k)}, \Gamma^{(k)}$. Composition with (10) yields the closed-form $\partial \mathcal{L}_K / \partial U$, $\partial \mathcal{L}_K / \partial V$.

Stationarity. At any $B_\star \in \arg\min_B \mathcal{L}_K$ (no exact-match assumption), $\nabla_B \mathcal{L}_K(B_\star) = 0$ pairs with $\Delta \in \mathbb{R}^{P \times C}$ via (29) to give

$$\sum_{k=1}^{K} w_k \, \langle \Gamma_0^{(k)}, \, L_k(\Delta) \rangle_F \;=\; 0, \qquad \Gamma_0^{(k)} := B_\star \Sigma_y^{(k)} B_\star^T - \Sigma^{\mathrm{ref},(k)}; \tag{30}$$

i.e. $\Gamma_0 := (\Gamma_0^{(k)})_k$ is orthogonal (under $\sum_k w_k \langle \cdot, \cdot \rangle_F$) to $\operatorname{im} L_{B_\star}$.

Hessian. Expand $\widehat{\Sigma}^{(k)}(B_\star + \Delta) - \Sigma^{\mathrm{ref},(k)} = \Gamma_0^{(k)} + L_k(\Delta) + \Delta \Sigma_y^{(k)} \Delta^T$, square, sum with weights w_k, and group by order in Δ:

$$\mathcal{L}_K(B_\star + \Delta) - \mathcal{L}_K(B_\star) \;=\; \underbrace{2 \sum_k w_k \langle \Gamma_0^{(k)}, L_k(\Delta) \rangle_F}_{=\,0 \text{ by (30)}} \;+\; H[\Delta] \;+\; 2R_3(\Delta) + R_4(\Delta),$$

with $H[\Delta] := \sum_k w_k \big(\|L_k(\Delta)\|_F^2 + 2\langle \Gamma_0^{(k)}, \Delta \Sigma_y^{(k)} \Delta^T \rangle_F \big)$ the quadratic-in-Δ contribution, $R_3(\Delta) := \sum_k w_k \langle L_k(\Delta), \Delta \Sigma_y^{(k)} \Delta^T \rangle_F$ purely cubic, and $R_4(\Delta) := \sum_k w_k \|\Delta \Sigma_y^{(k)} \Delta^T\|_F^2$ purely quartic. The linear-in-Δ contribution vanishes by (30), identifying the second-order term with (17). Under $\varepsilon_K^* = 0$, $\Gamma_0^{(k)} = 0$ for all k, H collapses to $\sum_k w_k \|L_k(\Delta)\|_F^2$, and rank H = rank L.

Proof of Lemma 9

The bound $D_\parallel$. Two sources of $D_\parallel$ are available. Under (15), Proposition 7 (3) gives $D_\parallel = c\rho + \|B(A_0) - B_\star\|_F$ along the whole admissible trajectory. When $\ker L_{B_\star} = \{0\}$ one may take $D_\parallel = 0$.

Set $\Delta := B - B_\star$ and split orthogonally $\Delta = \Delta_\perp + \Delta_\parallel$. By (b), the second-order Taylor expansion at $B_\star$ gives

$$\mathcal{L}_K(B) - \mathcal{L}_K(B_\star) \;\geq\; h \, \|\Delta_\perp\|_F^2 \;-\; \tfrac{M_3}{6} \, \|\Delta\|_F^3,$$

where M_3 is a constant with $|2R_3(\Delta) + R_4(\Delta)| \leq \frac{M_3}{6} \|\Delta\|_F^3$ on $\|\Delta\|_F \leq r_0$. Bounding $\|\Delta\|_F^3 \leq r_0(\|\Delta_\perp\|_F^2 + \|\Delta_\parallel\|_F^2)$ on the same ball and choosing $r_0 := h/M_3$ gives

$$\mathcal{L}_K(B) - \varepsilon_K^* \;\geq\; \tfrac{h}{2} \, \|\Delta_\perp\|_F^2 \;-\; \tfrac{h}{6} \, \|\Delta_\parallel\|_F^2. \tag{31}$$

Let $B \in \mathcal{B}(B_\star, \delta_B/2)$ with $\|\Delta_\perp\|_F \geq \delta_B/2$. Orthogonality gives $\|\Delta_\parallel\|_F^2 = \|\Delta\|_F^2 - \|\Delta_\perp\|_F^2 \leq 0$, so $\Delta_\parallel = 0$ and $\|\Delta_\perp\|_F = \delta_B/2$, and (31) gives $\mathcal{L}_K(B) - \varepsilon_K^* \geq h\delta_B^2/8 \geq \alpha^\star$. By contraposition every $B \in \mathcal{B}(B_\star, \delta_B/2)$ with $\mathcal{L}_K(B) - \varepsilon_K^* < \alpha^\star$ satisfies $\|\Delta_\perp\|_F < \delta_B/2$, which is the stated inclusion. That set is open, because $\mathcal{L}_K$ is continuous and the ball is open, and it contains $B_\star$, where the excess loss is zero. It contains no other element of $\mathcal{S}_K$, because by Proposition 4 every other element is at Frobenius distance at least δ_B from $B_\star$ and therefore outside the ball.

The same floor, evaluated on the sphere $\|\Delta\|_F = \delta_B/2$, is the second assertion of the lemma and is what Corollary 10 uses. There $\|\Delta_\parallel\|_F \le D_\parallel$ by hypothesis and $\|\Delta_\perp\|_F^2 = \delta_B^2/4 - \|\Delta_\parallel\|_F^2$, and the right-hand side of (31) decreases in $\|\Delta_\parallel\|_F^2$, so

$$\mathcal{L}_K(B) - \varepsilon_K^* \;\ge\; \frac{h}{2}\Big(\frac{\delta_B^2}{4} - D_\parallel^2\Big) \;-\; \frac{h}{6}\,D_\parallel^2 \;=\; \frac{h\,\delta_B^2}{8} \;-\; \frac{2h\,D_\parallel^2}{3} \;=\; \alpha^\star \tag{32}$$

for every B with $\|B - B_\star\|_F = \delta_B/2$. The two contributions $h/2$ and $h/6$ both multiply $D_\parallel^2$ and add to $2h/3$, which is the coefficient in (18). Under $D_\parallel < \delta_B/\sqrt{12}$ the threshold is positive, since then $\alpha^\star > h\delta_B^2/8 - h\delta_B^2/18 = 5h\delta_B^2/72$.

The bound $\|\Delta_\parallel\|_F \le D_\parallel$ has two sources. Under the warm-start hypothesis (15), Proposition 7 (3) supplies $D_\parallel = c\rho + \|B(A_0) - B_\star\|_F$ along the whole admissible trajectory. When $\ker L_{B_\star} = \{0\}$ the parallel component vanishes, so $D_\parallel = 0$ may be taken, $\alpha^\star = h\delta_B^2/8$, and the warm-start hypothesis is not needed.

Verification that $r_0 \ge \delta_B/2$ on the polynomial structure of $\mathcal{L}_K$. The remainder bound $|2R_3(\Delta) + R_4(\Delta)| \le \frac{M_3}{6}\|\Delta\|_F^3$ above is qualitative; we make M_3 explicit using that $\mathcal{L}_K$ is a degree-4 polynomial in B. From the proof of Theorem 8 (no exact-match assumption used),

$$\mathcal{L}_K(B_\star + \Delta) - \mathcal{L}_K(B_\star) \;=\; H(\Delta) + 2R_3(\Delta) + R_4(\Delta),$$

with

$$\begin{aligned} H(\Delta) &:= \sum_k w_k\big(\|L_k(\Delta)\|_F^2 + 2\langle \Gamma_0^{(k)},\, \Delta\Sigma_y^{(k)}\Delta^T\rangle_F\big),\\ R_3(\Delta) &:= \sum_k w_k\,\langle L_k(\Delta),\, \Delta\Sigma_y^{(k)}\Delta^T\rangle_F,\\ R_4(\Delta) &:= \sum_k w_k\,\|\Delta\Sigma_y^{(k)}\Delta^T\|_F^2. \end{aligned}$$

H is the second-order term of the expansion, hence $H(\Delta) \ge h\|\Delta_\perp\|_F^2$ by (b); R_3 is purely cubic and R_4 purely quartic in Δ. The bounds on R_3, R_4 below depend only on L_k and $\Delta\Sigma_y^{(k)}\Delta^T$, not on $\Gamma_0^{(k)}$, so the M_3 derivation is identical in both regimes. Submultiplicativity gives

$$\|L_k(\Delta)\|_F \le 2\|B_\star\|_{\mathrm{op}}\,\|\Sigma_y^{(k)}\|_{\mathrm{op}}\,\|\Delta\|_F, \qquad \|\Delta\Sigma_y^{(k)}\Delta^T\|_F \le \|\Sigma_y^{(k)}\|_{\mathrm{op}}\,\|\Delta\|_F^2,$$

and Cauchy–Schwarz on R_3 then yields, with $S := \sum_k w_k\|\Sigma_y^{(k)}\|_{\mathrm{op}}^2$,

$$2|R_3(\Delta)| \le 4\|B_\star\|_{\mathrm{op}}\,S\,\|\Delta\|_F^3, \qquad |R_4(\Delta)| \le S\,\|\Delta\|_F^4.$$

On the ball $\|\Delta\|_F \le r_0$,

$$\big|2R_3(\Delta) + R_4(\Delta)\big| \;\le\; \big(4\|B_\star\|_{\mathrm{op}} + r_0\big)\,S\,\|\Delta\|_F^3 \;=:\; \tfrac{M_3(r_0)}{6}\,\|\Delta\|_F^3,$$

which identifies $M_3(r_0) = \big(24\|B_\star\|_{\mathrm{op}} + 6r_0\big)\,S$.

Substituting into $r_0 = h/M_3(r_0)$ gives the quadratic $6S\,r_0^2 + 24\|B_\star\|_{\mathrm{op}}\,S\,r_0 - h = 0$, with positive root

$$r_0 \;=\; \frac{\sqrt{144\|B_\star\|_{\mathrm{op}}^2\,S^2 + 6Sh} \;-\; 12\|B_\star\|_{\mathrm{op}}\,S}{6S}.$$

Squaring, $r_0 \ge \delta_B/2$ is equivalent to (28). Under condition (b), (31) is valid throughout $\mathcal{B}(B_\star, \delta_B/2)$.

PROOF OF COROLLARY 10

Gradient flow gives $\frac{d}{dt}\mathcal{L}(A_t) = -\|\nabla\mathcal{L}(A_t)\|_F^2 \le 0$, so for every $t \ge t_0$, $\mathcal{L}_K(B_t) - \varepsilon_K^* \le \mathcal{L}_K(B_{t_0}) - \varepsilon_K^* < \alpha^\star$, the last inequality by condition (c).

The map $t \mapsto B_t = \Pi A_t^\dagger$ is continuous, because $t \mapsto A_t$ is continuous and $A \mapsto \Pi A^\dagger$ is continuous on $\mathcal{F}_A$, on which A has full row rank. Suppose that $\|B_{t_2} - B_\star\|_F \ge \delta_B/2$ for some $t_2 > t_0$. Condition (c) gives $\|B_{t_0} - B_\star\|_F < \delta_B/2$, so by continuity there is a first time $t_1 \in (t_0, t_2]$ with $\|B_{t_1} - B_\star\|_F = \delta_B/2$. Applying (32) at B_{t_1} gives $\mathcal{L}_K(B_{t_1}) - \varepsilon_K^* \ge \alpha^\star$, which contradicts the first display of this proof. Hence $\|B_t - B_\star\|_F < \delta_B/2$ for all $t \ge t_0$, that is $B_t \in \mathcal{B}(B_\star, \delta_B/2)$. By Proposition 4 no other element of $\mathcal{S}_K$ lies in that ball, so the sign class selected at t_0 is kept for all later times.

D SYNTHETIC-DATA EXPERIMENTS

D.1 SINGLE-PIXEL CALIBRATION: LORA PRIOR VS FREE A

This subsection supports the numerical illustration of Fig. 2 in the main text: a small calibration instance on which we train two models from the same warm-start A_0 under the same cov-matching loss, and read out the operator-recovery error and held-out reconstruction PSNR.

The instance has image dimension $N = 28^2 = 784$ and channel count $C = 64$. The companion operator $\Pi \in \mathbb{R}^{P\times N}$ is a 4×4 average-pooling map, so $P = 49$. The true operator is $A_\star = A_0 + U_\star V_\star^\top$, where A_0 is formed by stacking the row-normalised Π above $C - P$ further rows whose entries are independent $N(0, 1/N)$ variables, so that $\mathrm{row}(\Pi) \subseteq \mathrm{row}(A_0)$. The planted correction is $U_\star V_\star^\top = JA_0$ with J of rank $r_\star = 4$ and $\|J\|_{\mathrm{op}} = \rho_\star = 0.10$, so that $A_\star = (I_C + J)A_0$ and $\mathrm{row}(A_\star) = \mathrm{row}(A_0)$; by §A.3 the references are then exactly attainable, $\varepsilon_K^* = 0$.

Training is at the population level: we draw $K = 10$ block-constant source covariances $\{\Lambda_k\}_{k=1}^K$, each $\Lambda_k \in \mathbb{R}^{P\times P}$ sampled independently with spectral-concentration parameter $\beta = 2.0$, and an in-row-space source model $x = \Pi^\dagger s + \sigma_z \eta$ with isotropic noise $\sigma_z = 0.05$. Each Λ_k is $\mathrm{diag}(\ell_1^{(k)}, \dots, \ell_P^{(k)})$ with $\ell_i^{(k)} = \exp(\beta z_{ik})$, where the z_{ik} are independent $N(0,1)$ variables and $\beta = 2.0$ is the spectral-concentration parameter. The sources are $s \sim N(0, \Lambda_k)$ and the noise is $\eta \sim N(0, I_N)$, independent of s, so that $\Sigma_x^{(k)} = \Pi^\dagger \Lambda_k (\Pi^\dagger)^\top + \sigma_z^2 I_N$. Each Λ_k is diagonal and, the pooling patches being disjoint, $\Pi\Pi^\top$ is diagonal as well, so each D_k is diagonal and the hypothesis of Proposition 3 holds in the standard basis. From these we form $\Sigma_y^{(k)} = A_\star \Sigma_x^{(k)} A_\star^\top$ and $\Sigma^{\mathrm{ref},(k)} = \Pi \Sigma_x^{(k)} \Pi^\top$ for $k = 1, \dots, K$, the latter equal to $\Lambda_k + \sigma_z^2 \Pi\Pi^\top$ since $\Pi\Pi^\dagger = I_P$, giving $K \cdot P(P+1)/2 \approx 1.2 \times 10^4$ independent constraints. These are the closed-form population covariances of the model above, so training draws no samples. The Gaussian law of s and η is used only to draw the held-out set, and hence enters only the PSNR readout below.

Two models are trained, both warm-started at A_0. The LoRA model is $A(U, V) = A_0 + UV^\top$ with $U \in \mathbb{R}^{C\times r}$, $V \in \mathbb{R}^{N\times r}$, $r = r_\star = 4$, and U, V initialised at scale 10^{-3}, giving $r(C + N) = 3392$ trainable parameters. The free-A baseline uses A_0 as the initial value of an unconstrained trainable parameter, giving $CN = 50176$ trainable parameters — well above the $\approx 1.2 \times 10^4$ constraints, so the matching set in unstructured A-space is a high-dimensional manifold rather than a discrete set. Both models minimise the same loss

$$\mathcal{L}(A) \;=\; \sum_{k=1}^{K} \left\| \Pi A^\dagger \Sigma_y^{(k)} (A^\dagger)^\top \Pi^\top \;-\; \Sigma^{\mathrm{ref},(k)} \right\|_F^2,$$

that is (5) with $w_k = 1$, optimised with Adam at learning rate 2×10^{-3}, cosine-annealed to 2×10^{-5}, for 3000 iterations; the two runs share the same random seed.

Held-out reconstruction is evaluated on 1000 samples (100 per Λ_k) drawn from the same in-row-space distribution as the training covariances. We use Tikhonov inversion $\hat{x} = (A^\top A + \lambda I)^{-1} A^\top y$ with $\lambda = 10^{-3}$ and $y = A_\star x$, and report PSNR with the global dynamic range of x.

Numerical readout. After 3000 iterations both models have driven the training loss below 10^{-7}, but they sit at very different points. The free-A optimum has $\|A - A_\star\|_F/\|A_\star\|_F \approx 0.20$ and $\|B - B_\star\|_F/\|B_\star\|_F \approx 0.037$ with $B := \Pi A^\dagger$, against 0.018 and 0.014 respectively for LoRA. Held-out PSNR means: A_0 62.37 dB, free A 53.87 dB, LoRA 63.89 dB, oracle $A_\star$ 65.22 dB. The free-A trajectory escapes the warm-start neighbourhood and lands on an arbitrary element of the high-dimensional matching manifold, whereas the LoRA parameterisation restricts training to a neighbourhood of the target discrete element of $\mathcal{S}_K$ predicted by claim (I).

Rank ablation. Keeping the instance, the warm start and the optimiser settings fixed, we vary the training rank over $r \in \{1, 2, 4, 8, 16, 32\}$ and record the final training loss, the held-out PSNR, the relative operator error and the operator norm of the correction. Table 2 reports them. The column $r = 4$ is the run of Fig. 2 and reproduces the readout above.

r	1	2	4	8	16	32
final loss	$5.6{\cdot}10^{1}$	$2.8{\cdot}10^{1}$	$1.0{\cdot}10^{-7}$	$1.1{\cdot}10^{-7}$	$4.1{\cdot}10^{-7}$	$2.4{\cdot}10^{-7}$
held-out PSNR (dB)	61.79	63.24	63.89	64.10	64.12	63.60
$\|A - A_\star\|_F/\|A_\star\|_F$	0.0273	0.0197	0.0179	0.0162	0.0162	0.0243
final $\|UV^\top\|_{\text{op}}$	0.165	0.113	0.107	0.097	0.111	0.163
$\max_t \|U_t V_t^\top\|_{\text{op}}$	0.165	0.114	0.108	0.107	0.119	0.172

Table 2: Rank ablation on the instance of Fig. 2, one run per rank, with the same seed 0 as that figure. The last two rows are the operator norm of the correction at the last iteration and its maximum over the logged iterations of the run; for reference, $\sigma_{\min}(A_0) = 0.769$ and the admissibility ceiling of Proposition 7 (1) is $\sigma_{\min}(A_0)/2 = 0.385$.

The planted rank is $r_\star = 4$. Below it the final loss stays about eight orders of magnitude above the value reached at $r_\star$, and at $r = 1$ the held-out PSNR ends below the warm-start value 62.37 dB. From $r = 4$ on the held-out PSNR varies by 0.52 dB and the relative operator error by 0.0081, so the held-out PSNR and the relative operator error saturate above the planted rank. In every run the correction remains admissible: the largest value of $\|U_t V_t^\top\|_{\text{op}}$ over all six runs and all logged iterations is 0.172, below the ceiling 0.385.

The ablation above sits at one warm start, and the conclusion about the rank does not depend on that choice. Table 3 below runs $\sigma_{\min}(A_0) \in \{10^{-2}, 10^{-3}\}$ at $\sigma_{\max}(A_0) = 1$ against three aggregators, and the warm start of the present subsection is built differently and is well conditioned, with $\sigma_{\min}(A_0) = 0.769$. At both ends of that range the profile is the same: below $r_\star$ there is no usable solution, and above it there is no gain.

The scope of the numerical work is the following. Each cell of Table 2 and of Table 3 is one run at one seed. In the instance of this subsection the noise level σ_z and the size $\rho_\star$ of the planted mismatch are each held at one value, and the comparison is against an unconstrained A, with no third regularised baseline. The spread reported in Table 4 is a maximum over seed pairs rather than an error bar.

We further report two quantitative probes on a separate synthetic instance, targeting analytical bounds rather than the matching-set picture of §D.1: a rank scan validating the rank-monotone amplification bound of Proposition 6, and a trajectory-anchor demonstration validating the warm-start B-distance bound of Proposition 7 (1).

D.2 Setup

The synthetic instance has $C = 60$, $N = 1024$, and $P \in \{8, 16, 32\}$. The baseline A_0 is drawn with prescribed extreme singular values $\sigma_{\min}(A_0) \in \{10^{-2}, 10^{-3}\}$ and $\sigma_{\max}(A_0) = 1$ on a log-spaced spectrum. A planted target $A_{\text{target}} = A_0 + U_{\text{target}} V_{\text{target}}^T$ has rank $r_{\text{target}} = 4$ and $\|U_{\text{target}} V_{\text{target}}^T\|_{\text{op}} = 0.3\,\sigma_{\min}(A_0)$. The aggregator is $\Pi = [I_P, 0_{P\times(N-P)}]$, so $\|\Pi\|_{\text{op}} = 1$, and $B_{\text{target}} := \Pi A_{\text{target}}^\dagger$. The references are $K = 3$: one symmetric covariance $\Sigma_y^{(1)}$ and two non-symmetric lag-τ cross-covariances. They are constructed so that the whitened operator $W := L^{-1} B_{\text{target}}$ ($LL^T = B_{\text{target}} B_{\text{target}}^T$ Cholesky) simultaneously diagonalises $\{\Sigma_y^{(k)}\}$, with diagonal entries drawn from $\text{Unif}(0.5, 2.0)$ and free directions on the $C-P$ complement from $\text{Unif}(10^{-2}, 10^{-1})$;

$\Sigma^{\mathrm{ref},(k)} := B_{\mathrm{target}} \Sigma_y^{(k)} B_{\mathrm{target}}^T$ analytically. By construction B_{target} and its 2^P sign-flip equivalents fill $\mathcal{S}_K$.

The training parameterisation is $A = A_0 + UV^T$ with trainable $U \in \mathbb{R}^{C\times r}, V \in \mathbb{R}^{N\times r}$. Optimiser: Adam with learning rate 10^{-4} and initial scale 10^{-5} for $T = 6000$ full-batch steps, in float64. The reported loss is the unnormalised $\mathcal{L}_K$; optimisation runs on a normalised version (divided by $\sum_k w_k \|\Sigma^{\mathrm{ref},(k)}\|_F^2$, $w_k = 1/K$) to keep gradient magnitudes comparable across $\sigma_{\min}(A_0)$. Two protocols are reported: (a) a rank scan with $r \in \{1, 2, 4, 6, 8, 12, 16, 20\}$ at one seed per cell, across all six $(P, \sigma_{\min}(A_0))$ configurations (Proposition 6); (b) a trajectory-anchor demonstration at $r = r_{\mathrm{target}} = 4$, $P = 8$, with three seeds for $\sigma_{\min}(A_0) = 10^{-2}$ and five seeds for $\sigma_{\min}(A_0) = 10^{-3}$, the latter with explicit per-step projection enforcing $\|UV^T\|_{\mathrm{op}} \le \sigma_0/2$ (Proposition 7).

D.3 Results

Rank scan. For each $(P, \sigma_{\min}(A_0))$ cell, define $\widehat{\mathcal{L}^\star}(r) := \min_{t\ge 0.75T} \mathcal{L}_K(B_t)$ (minimum over the last 25% of training steps, robust to optimisation oscillation) and $r_{\min} := \arg\min_r \widehat{\mathcal{L}^\star}(r)$. Table 3 reports $r_{\min}$ per cell.

P	$\sigma_{\min}(A_0)$	$r_{\min}$	$\widehat{\mathcal{L}^\star}(r_{\min})$
8	10^{-2}	20^*	6.8×10^{-4}
8	10^{-3}	8	1.8×10^{1}
16	10^{-2}	20^*	3.5×10^{-3}
16	10^{-3}	8	9.2×10^{1}
32	10^{-2}	12	2.1×10^{-2}
32	10^{-3}	12	2.4×10^{2}

Table 3: Argmin of the rank scan $r_{\min} := \arg\min_{r\in\{1,2,4,6,8,12,16,20\}} \widehat{\mathcal{L}^\star}(r)$, single seed per cell. A starred argmin marks a flat-tail cell: $\widehat{\mathcal{L}^\star}(r)$ varies by less than a factor of two across $r \in \{4, 6, \ldots, 20\}$ at the optimisation noise floor, and the in-grid argmin is dominated by training oscillation rather than by $\bar\kappa(r)$.

In each $\sigma_{\min}(A_0) = 10^{-3}$ cell the scan exhibits an interior minimum at $r_{\min} \in \{8, 12\}$, with a sharp decrease for $r \le r_{\min}$ and a rebound for $r > r_{\min}$. The rebound coincides numerically with the entry into the inadmissible region $r\rho^2 \gtrsim \sigma_{\min}(A_0)$, where Proposition 6 predicts inflated back-prop amplification: e.g., at $(P, \sigma_{\min}(A_0)) = (16, 10^{-3})$, $\widehat{\mathcal{L}^\star}(r) \approx (4.4{\cdot}10^5, 6.5{\cdot}10^3, 2.0{\cdot}10^2, 1.0{\cdot}10^2, 9.2{\cdot}10^1, 1.1{\cdot}10^3, 4.1{\cdot}10^4, 5.5{\cdot}10^4)$ at $r = (1, 2, 4, 6, 8, 12, 16, 20)$, with the rebound at $r = 12$ matching the operator-norm crossing $\|U_t V_t^T\|_{\mathrm{op}}^{\max} \approx 3.4 \times 10^{-3} > \sigma_{\min}(A_0)$. For $\sigma_{\min}(A_0) = 10^{-2}$ with $P \in \{8, 16\}$, admissibility does not bind within the scan range and $\widehat{\mathcal{L}^\star}(r)$ saturates at the optimisation noise floor for $r \ge 4$; the in-grid argmin is determined by training oscillation. At the largest P with $\sigma_{\min}(A_0) = 10^{-2}$ an interior minimum at $r_{\min} = 12$ does emerge.

Trajectory-anchor demonstration. Two trained operators $B_\infty^{(i)}, B_\infty^{(j)}$ from different initialisations under the same A_0 and $\{\Sigma_y^{(k)}\}$ satisfy

$$\|B_\infty^{(i)} - B_\infty^{(j)}\|_F \le 2c\rho + 2\|B(A_0) - B_\star\|_F \tag{33}$$

under the warm-start admissibility hypothesis, where $B_\star$ is the common nearest element of $\mathcal{S}_K$ to $B(A_0)$. We evaluate this at $C = 60$, $P = 8$, $K = 3$, $r = r_{\mathrm{target}} = 4$ on two configurations: $\sigma_{\min}(A_0) = 10^{-2}$ (3 seeds, post-hoc on the rank scan cache without explicit projection) and $\sigma_{\min}(A_0) = 10^{-3}$ (5 seeds, fresh training with per-step $\|UV^T\|_{\mathrm{op}}$ projection at the admissibility ceiling $\rho_{\mathrm{admiss}} := \sigma_0/2 = 5 \times 10^{-4}$ from Proposition 7 (1)). $B(A_0)$ and $B_\star$ are computed analytically: the planted B_{target} and its 2^P sign-flip equivalents fill $\mathcal{S}_K$ in this construction, and $B_\star$ is the equivalent nearest to $B(A_0)$.

$\sigma_{\min}(A_0)$	$\rho_{\max}$	admissible?	$\|B(A_0) - B_\star\|_F$	(a)	(b)	(a)/(b)
10^{-2}	9.8×10^{-4}	yes	1.43	1.27	317	4.0×10^{-3}
10^{-3}	5.0×10^{-4}	yes (enforced)	12.98	9.64	1.60×10^4	6.0×10^{-4}

Table 4: Pairwise B_∞-distance demonstration of (33) at $C = 60$, $P = 8$, $K = 3$, $r = r_{\text{target}} = 4$. $\rho_{\max} := \max_t \|U_t V_t^T\|_{\text{op}}$ (here t is the training-step index, see Proposition 7); empirical (a) is $\max_{i,j} \|B_\infty^{(i)} - B_\infty^{(j)}\|_F$ over seed pairs; predicted (b) is $2c\rho_{\max} + 2\|B(A_0) - B_\star\|_F$ from Proposition 7 (1). The bound holds in both rows, the second by per-step $\|UV^T\|_{\text{op}}$ projection enforced at $\sigma_0/2$. The slack between (a) and (b) traces to the $\sqrt{r}$ Frobenius factor in c and the $\sigma_{\min}(A(s)) \geq \sigma_0/2$ Weyl simplification used in the bound.

D.4 Discussion

The rank scan shows $\mathcal{L}^\star(r)$ interior-minimised in the regime where the rank-binding ceiling $r\rho^2 < \sigma_{\min}(A_0)$ from Proposition 6 (a different ceiling from the op-norm admissibility $\rho \leq \sigma_0/2$ of Proposition 7) binds within the scanned r range: the loss profile is non-monotone with $r_{\min}$ strictly interior, and the rebound at $r > r_{\min}$ coincides with violation of this rank-binding ceiling, matching Proposition 6. When admissibility does not bind in-range, the loss saturates and the interior-optimum behaviour is consistent with but not effectively tested by the scan.

The trajectory-anchor demonstration verifies the Frobenius bound (33) of Proposition 7 (1) at two $\sigma_{\min}(A_0)$ values; the bound is loose by a factor of ≈ 250 ($\sigma = 10^{-2}$) to ≈ 1700 ($\sigma = 10^{-3}$), with slack scaling as $\sqrt{r}$ from $\|UV^T\|_F \leq \sqrt{r}\,\|UV^T\|_{\text{op}}$ and as σ_0^{-2} from the $\sigma_{\min}(A(s)) \geq \sigma_0/2$ Weyl simplification used in the bound. In the present setting $C > P$ and so $\dim \ker L_{B_\star} \geq P(C-P) > 0$; the δ_B-ball confinement in Lemma 9 requires the a-priori bound $D_\parallel$ on the $\ker L_{B_\star}$ component of $B - B_\star$ supplied by (16) (Proposition 7 (3)), and the trajectory-anchor demonstration is a direct numerical test of this $D_\parallel$ bound.

D.5 Rank of the stacked differential against K

The dimension count after Proposition 2 gives $\operatorname{rank} L_B \leq \sum_k n_k$, with $n_k = P(P+1)/2$ for a symmetric reference and $n_k = P^2$ for a non-symmetric one. We check whether that bound is the only constraint on the rank, that is, whether the rank reaches PC at the first K with $\sum_k n_k \geq PC$.

The evaluation point is a generic $B \in \mathbb{R}^{P\times C}$ with independent standard normal entries, and the reference covariances are drawn at random as well. This point is not a solution of any matching system, so it is not the point $B_\star$ at which Proposition 2 states its rank condition, and the $P(C-P)$ kernel directions of (26) are absent here. The quantity in the table is the numerical rank of the stacked L_B computed with the absolute tolerance 10^{-8} on the singular values.

(P, C)	reference	n_k	PC	K=1	K=2	K=3	K=4
$(2, 4)$	symmetric	3	8	3	6	8	8
$(2, 4)$	non-symmetric	4	8	4	8	8	8
$(3, 5)$	symmetric	6	15	6	12	15	15
$(3, 5)$	non-symmetric	9	15	9	15	15	15

Table 5: Numerical rank of the stacked differential $L_B = (L_1, \ldots, L_K)$ at a generic $B \in \mathbb{R}^{P\times C}$, at tolerance 10^{-8}.

In each row the rank equals $\min(Kn_k, PC)$, so the first K at which it attains PC is $\lceil PC/n_k \rceil$, which is the first K with $\sum_k n_k \geq PC$. The counting bound is therefore the only constraint that binds at these generic evaluation points.

E Hypothesis ledger

Table 6 lists every hypothesis used in the paper, where it is stated, which results use it, what fails without it, and whether it can be checked from $(A_0, \Pi, \Sigma_y^{(k)}, \Sigma^{\mathrm{ref},(k)})$ before any training step (column "Before").

Table 6: Hypothesis ledger. Column three lists the results whose proofs use the hypothesis, directly or through another result they invoke. Column five records whether the hypothesis can be checked from $(A_0, \Pi, \Sigma_y^{(k)}, \Sigma^{\mathrm{ref},(k)})$ before any training step.

Hypothesis	**Stated at**	**Used by**	**What fails without it**	**Before**
A, hence A_0, of full row rank	notation table of main text §3; Lemma 5; Proposition 7	Lemma 5, Propositions 6 and 7, Corollary 10, Remark 11	$A \mapsto A^\dagger$ is discontinuous where the rank drops, so (10) does not apply there	yes
$\Sigma_y^{(k)} \succ 0$, $\Sigma^{\mathrm{ref},(k)} \succ 0$, $w_k > 0$	notation table of main text §3; (5)	Propositions 1 to 4, Theorem 8	no whitening in Propositions 1 and 3; $\delta_B = 0$ in (8)	yes
Π known and fixed, $P < N$	main text §1; notation table of §3	every result through $B := \Pi A^\dagger$; quantitatively in (13) and (19)	B is not defined as a function of A	yes
$\varepsilon_K^* = 0$	main text §4.1	Propositions 2 to 4 and Proposition 7 (2), (3); Lemma 9 and Corollary 10 through condition (a). Not used by Lemma 5, Proposition 6, Proposition 7 (1), Theorem 8	$\mathcal{S}_K$ is no longer the solution set of the matching equations, so the sign-class decomposition and δ_B lose their content. Lemma 9 and Corollary 10 use it only through δ_B and condition (a); their own statements are in terms of the excess loss and of a general bound $D_\parallel$. Row-space inclusion is a sufficient condition for it, see §A.3	yes
counting condition $\sum_k n_k \geq PC$	Proposition 2, as a consequence of the rank condition	no later result uses it; it constrains the design of (P, C, K) and the choice between symmetric and lagged references	fewer equations than the PC unknowns. It is not necessary for discreteness, see §A.2	yes
rank condition $\operatorname{rank} L_{B_\star} = PC$	main text §4.1; Proposition 2	sufficiency in Proposition 2	the differential $L_{B_\star}$ is no longer injective, so discreteness does not follow from a first-order argument. Under $\varepsilon_K^* = 0$, (6) and (7) it cannot hold at all when $P < C$, by (26), although Proposition 3 still gives a discrete matching set	yes
$\lambda_1 > \cdots > \lambda_{\min(P,C)} > 0$	Proposition 3	Propositions 3 and 4; Lemma 9 and Corollary 10 through condition (a)	positivity: no whitening and $\delta_B = 0$. Strictness: Q is no longer fixed by Σ^{ref}, so the sign action of Proposition 3 is not canonical	yes
each D_k diagonal	Proposition 3; condition (a) of §C above	as in the row above	joint diagonalisation becomes joint similarity, and the finiteness of $\mathcal{S}_K$ is lost	yes
$\boldsymbol{\rho}_i$ pairwise distinct	Proposition 3	as in the row above	two equal profiles leave an $O(2)$ block in the residual symmetry group, so a continuum of loss-equivalent B appears	yes

Hypothesis	Stated at	Used by	What fails without it	Before
(6) and (7), for $P < C$	main text §4.1	Proposition 3 when $P < C$, and everything that goes through condition (a) in that regime	the reduction does not close and the matching set can be a continuum; the remark in §A.1 gives such an instance	yes
shared column norm ρ	main text §4.2	Proposition 6 only	$\|UV^T\|_{\text{op}} \le r\rho^2$ and $r_{\max}$ in (12) lose their meaning; no other result changes	yes
admissibility $\|UV^T\|_{\text{op}} \le \rho \le \sigma_0/2$	Proposition 7 (1)	Proposition 7; Lemma 9 and Corollary 10 through (15) and (16)	Weyl's inequality gives no floor on $\sigma_{\min}(A(s))$, so the constant (13) is unavailable	yes
a bound $D_\parallel$ on the $\ker L_{B_\star}$ component of $B - B_\star$	Lemma 9	Lemma 9 and Corollary 10	(18) has no finite value. Proposition 7 (3) supplies $D_\parallel = c\rho + \|B(A_0) - B_\star\|_F$ under (15), and $D_\parallel = 0$ may be taken when $\ker L_{B_\star} = \{0\}$	yes
design target κ_0	main text §4.2, (11)	$\mathcal{F}_A$; (12); the Lipschitz constant and (19) of Remark 11	no uniform Lipschitz constant on $\mathcal{F}_A$. This is a tuning parameter: any $\kappa_0 > 2/\sigma_{\min}(A_0)^2$ may be chosen	yes
condition (b), with h above (28)	§C above	Lemma 9 and hence Corollary 10	(31) is vacuous and the trust radius r_0 drops below $\delta_B/2$. Under $\varepsilon_K^* = 0$ the positivity on $\ker L_{B_\star}^\perp$ is a conclusion of Theorem 8, and only the quantitative threshold is exogenous	yes
condition (c)	§C above	Corollary 10 only	the sublevel set of Lemma 9 need not contain the trajectory. Remark 11 states an alternative that does not use it	no
unique nearest $B_\star$ to $B(A_0)$	Proposition 7 (2)	Proposition 7 (2), (3); Lemma 9; Corollary 10. Not used by Remark 11	under a tie between two elements of $\mathcal{S}_K$ the warm-start basin is not defined; the set of warm starts that produce a tie has measure zero, as main text §5 notes	yes