WIRTINGER FLOW FOR BLIND DECONVOLUTION FROM RANDOM INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the dynamics of Wirtinger flow (or gradient descent) for solving blind deconvolution from a random initialization. For bilinear measurements, we prove that the gradient method reaches an ϵ-accurate solution in $\mathcal{O}\big(\eta^{-1}\log m + \eta^{-1}\log(1/\epsilon)\big)$ iterations when $m \gtrsim K^3 \log^{11} m$. The analysis reveals two stages. During Stage I, the mismatch between the two vectors and their orthogonal components decay, allowing a weak common signal component to become dominant. Once the iterates enter the local region, they converge linearly during Stage II. We further extend the algorithm and the local analysis to blind deconvolution and demixing, where the update has a contraction coefficient independent of the source condition number. Numerical experiments demonstrate the main findings of our analysis.

1 INTRODUCTION

Blind deconvolution (BD) aims to recover two unknown vectors from bilinear measurements

$$y_j = \mathbf{b}_j^{\mathsf{H}} \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}} \mathbf{a}_j, \qquad 1 \le j \le m, \tag{1}$$

where $\mathbf{h}^\star, \mathbf{x}^\star \in \mathbb{C}^K$ are the unknown vectors, $\mathbf{a}_j$ is modeled to follow a complex Gaussian distribution, i.e., $\mathbf{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ for $1 \le j \le m$, and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]^{\mathsf{H}}$ is formed by the first K columns of a unitary m-point discrete Fourier transform matrix. This bilinear model is obtained by assuming the unknown convolutional signals belong to known low-dimensional subspaces of dimension K ($K < m$) and taking a discrete Fourier transform of their convolution. BD is a basic inverse problem in many scenarios, including imaging, communications, acoustics, etc. Representative applications include image deblurring (Kundur & Hatzinakos, 1996), blind channel identification and equalization (Hatzinakos, 1995), and speech dereverberation (Lin et al., 2007).

1.1 RELATED WORKS

Algorithmically, lifting (1) to the matrix variable $\mathbf{X} = \mathbf{h}\mathbf{x}^{\mathsf{H}}$ leads to convex nuclear-norm programs with exact-recovery guarantees (Ahmed et al., 2014). However, this method replaces two vectors by a matrix and can incur substantial memory and computational costs. This has motivated nonconvex methods that optimize the two vectors directly, including regularized gradient descent (Li et al., 2019), optimization on a quotient manifold (Huang & Hand, 2018), and Wirtinger flow (Ma et al., 2018). Existing guarantees for these fast first-order methods are local. A spectral initializer is constructed from the measurements to place the first iterate near the truth and inside an incoherent region. In this paper, we ask whether the Wirtinger flow proposed by Ma et al. (2018) can reach this local region from a completely data-independent random point, and if so, which dynamics govern the transient before local contraction begins.

Lifting also connects BD to low-rank matrix recovery. For matrix-sensing operators that satisfy the restricted isometry property (RIP), landscape analyses establish the absence of spurious local minima and the strict-saddle property (Bhojanapalli et al., 2016; Ge et al., 2017). These geometric results justify generic saddle-escaping algorithms, but do not by themselves characterize the transient or convergence rate of vanilla gradient descent from a random point. More recent work does study the iterates of gradient descent from a small random initialization in overparameterized matrix sensing (Li et al., 2018; Stöger & Soltanolkotabi, 2021). There the initial factor matrix has a norm close to

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.

zero, and the symmetric factorization $\mathbf{U}\mathbf{U}^{\mathsf{H}}$ allows the signal subspace to be learned progressively. For asymmetric factorizations, global convergence from a small initialization has been established for fully observed matrix factorization (Ye & Du, 2021), whereas the available matrix-sensing theory starts from an accurate initializer (Ma et al., 2021).

Our rank-one BD setting has three features that require a separate analysis. First, each measurement matrix is formed from a deterministic Fourier vector and a Gaussian vector. Controlling this structured ensemble requires Fourier incoherence in addition to estimation error. Second, BD uses the asymmetric factorization $\mathbf{h}\mathbf{x}^{\mathsf{H}}$ and initializes both vectors at unit norm. This combination produces dynamics beyond those of symmetric $\mathbf{U}\mathbf{U}^{\mathsf{H}}$ factorization started close to zero. Third, local asymmetric matrix-sensing results start from an accurate estimate, whereas we analyze the entire path from an independent random point to the local region. We address these features by deriving the population dynamics of the two BD vectors and then using a low-rank RIP to control the discrepancy between the population and finite-sample updates.

The closest analysis of random initialization is for phase retrieval (Chen et al., 2019). In that problem, the iterate is a single vector. Its projection onto the truth is the useful signal component, and the remainder is one orthogonal component. The analysis shows that their ratio grows until the iterate enters a region of linear convergence. BD instead involves two vectors and therefore has two correlations with the truth. Their average measures the signal shared by the two vectors, while their difference measures how far the two correlations are from one another. This difference is an additional error quantity that has to decay along with the two orthogonal components. This mismatch is invariant under the reciprocal rescaling of the two vectors that preserves their outer product. In addition, the update of each vector is divided by the squared norm of the other one. Thus the two vector norms and the signal and error components evolve together during the initial transient. The analysis must show both that these norms remain bounded away from zero and that the shared signal eventually grows. These coupled dynamics and the Fourier–Gaussian sensing structure are the main additional ingredients in the BD analysis.

Additionally, Ling & Strohmer (2019); Dong & Shi (2018; 2019) analyzed regularized or unregularized gradient descent for the joint blind deconvolution and demixing (BDD) problem. Both the theories and the experiments show that the linear-convergence rate deteriorates as the condition number measuring the imbalance among source strengths increases, and removing this dependence was left open. Fortunately, our experiments show that applying Chen et al. (2019)'s Wirtinger flow to BDD produces overlapped or parallel convergence curves, even with random initialization. This observation motivates our investigation of local convergence for BDD. Our curvature analysis establishes a local contraction coefficient independent of the source condition number.

Other related work studies sparse BD (Zhang et al., 2020; 2021), sphere-constrained formulations (Li & Bresler, 2019; Zhang et al., 2021), and multichannel observations (Li & Bresler, 2019; Qu et al., 2020). Several of these works prove convergence from random initializationt, but consider different models from (1).

1.2 Main Contributions

Our main contributions are summarized as follows.

- We establish a two-stage global convergence guarantee for Wirtinger flow from independent random initialization. The iterates reach the local region within $\mathcal{O}(\eta^{-1}\log m)$ steps and then converge linearly to the ground truth.
- We characterize the Stage-I mechanism by tracking the common signal component, the mismatch between the two vector correlations, the two orthogonal components, and the product of the two vector norms in the population dynamics. A uniform low-rank RIP for the Fourier–Gaussian sensing operator then transfers the population dynamics to the finite-sample iterates.
- We extend the local convergence analysis to BDD and show that the contraction coefficient of the update is independent of the source condition number.
- Numerical experiments illustrate the two-stage BD dynamics, the effects of the step size and sample size, and the condition-number-independent local convergence of BDD.

2 Randomly initialized Wirtinger flow for blind deconvolution

We work with the bilinear model (1) and minimize the empirical loss

$$f(\mathbf{h}, \mathbf{x}) = \sum_{j=1}^{m} \left| \mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j \right|^2 . \tag{2}$$

For every nonzero $c \in \mathbb{C}$, $(\mathbf{h}/\overline{c}, c\mathbf{x})$ produces the same observations as $(\mathbf{h}, \mathbf{x})$. Hence the solution to (2) is identifiable only up to the scaling ambiguity.

To minimize the empirical loss, we use a randomly initialized Wirtinger flow algorithm for BD. Denote the tth iterate by $\mathbf{z}^t := [\mathbf{h}^{t\mathsf{H}}, \mathbf{x}^{t\mathsf{H}}]^{\mathsf{H}}$ and write $\mathbf{z}^\star := [\mathbf{h}^{\star\mathsf{H}}, \mathbf{x}^{\star\mathsf{H}}]^{\mathsf{H}}$. Moreover, we use $f(\mathbf{z})$ to denote $f(\mathbf{h}, \mathbf{x})$. Let $\mathbb{S}_{\mathbb{C}}^{K-1} := \{\mathbf{u} \in \mathbb{C}^K : \|\mathbf{u}\|_2 = 1\}$ denote the complex unit sphere. We sample $\mathbf{h}^0$ and $\mathbf{x}^0$ independently and uniformly from $\mathbb{S}_{\mathbb{C}}^{K-1}$. Set $l_j := \mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j$. The Wirtinger gradients of (2) are

$$\nabla_{\mathbf{h}} f = \sum_{j=1}^{m} l_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}, \qquad \nabla_{\mathbf{x}} f = \sum_{j=1}^{m} \bar{l}_j \mathbf{a}_j \mathbf{b}_j^{\mathsf{H}} \mathbf{h}. \tag{3}$$

We apply the Wirtinger gradient update used in Ma et al. (2018), in which the update step is

$$\mathbf{h}^{t+1} = \mathbf{h}^t - \eta \frac{\nabla_{\mathbf{h}} f(\mathbf{z}^t)}{\|\mathbf{x}^t\|_2^2}, \qquad \mathbf{x}^{t+1} = \mathbf{x}^t - \eta \frac{\nabla_{\mathbf{x}} f(\mathbf{z}^t)}{\|\mathbf{h}^t\|_2^2}. \tag{4}$$

The denominators applied to the gradient terms in (4) make the update equivariant to complex rescaling and compensate for the curvature contributed by the other vector. Algorithm 1 summarizes the complete procedure.

Algorithm 1 Randomly initialized Wirtinger flow for blind deconvolution

Require: Measurements $\{(\mathbf{a}_j, \mathbf{b}_j, y_j)\}_{j=1}^m$, step size η, and iteration count T

1: Sample $\mathbf{h}^0$ and $\mathbf{x}^0$ independently and uniformly from $\mathbb{S}_{\mathbb{C}}^{K-1}$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Compute $\nabla_{\mathbf{h}} f(\mathbf{z}^t)$ and $\nabla_{\mathbf{x}} f(\mathbf{z}^t)$ using (3)
4: Update $\mathbf{h}^{t+1}$ and $\mathbf{x}^{t+1}$ via (4)
5: **end for**
6: **return** $\mathbf{h}^T, \mathbf{x}^T$

Figure 1 shows the behavior of the iterate of Algorithm 1. Each curve is an independent trial with $m = 10K$ and $\eta = 0.1$. The reconstruction error changes only slowly during an initial transient, and then decreases along an approximately straight line on the semilogarithmic scale. We refer to these two stages of the trajectory as Stage I and Stage II, respectively.

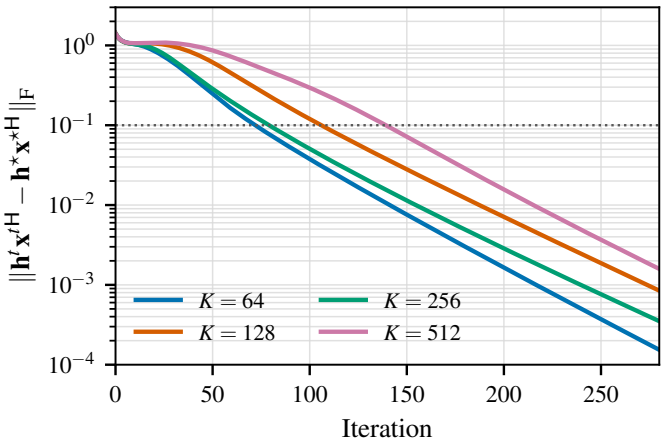


Figure 1: Two-stage convergence from random initialization.

3 Main Result: Two-Stage Convergence

Before stating the main result, we make two standard assumptions for the ground truth. First, without loss of generality, we assume $\|\mathbf{h}^\star\|_2 = \|\mathbf{x}^\star\|_2 = 1$. Second, following Ma et al. (2018), define the Fourier incoherence parameter μ with $1 \le \mu \le \sqrt{K}$ as the smallest number satisfying

$$\max_{1\le j\le m} |\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^\star| \le \frac{\mu}{\sqrt{m}} \|\mathbf{h}^\star\|_2 . \tag{5}$$

This condition quantifies the spectral flatness of the ground truth. A smaller μ indicates that the energy of $\mathbf{h}^\star$ is more evenly distributed across the Fourier basis, while a larger μ indicates that the energy is concentrated at a few Fourier frequencies. The incoherence condition (5) is standard in BD and low-rank recovery (Ahmed et al., 2014; Li et al., 2019; Ma et al., 2018), and is satisfied with high probability when $\mathbf{h}^\star$ is sampled uniformly from the complex unit sphere.

Moreover, to take the scaling ambiguity into account, we use the quotient distance to measure the optimization error

$$\operatorname{dist}(\mathbf{z}, \mathbf{z}^\star) := \min_{c\in\mathbb{C}\setminus\{0\}} \sqrt{\|\mathbf{h}/\overline{c} - \mathbf{h}^\star\|_2^2 + \|c\mathbf{x} - \mathbf{x}^\star\|_2^2}. \tag{6}$$

Theorem 1 (Global convergence from random initialization)**.** *There are numerical constants $c, C, C_0 > 0$ such that the following holds. Suppose $m \ge CK^3 \log^{11} m$ and $0 < \eta \le c/\log^5 m$. Then, with probability at least $1 - \mathcal{O}(m^{-10} + me^{-cK})$, there exists an index $T_\gamma \le C_0\eta^{-1}\log m$, with $\gamma = c/(\sqrt{K}\log^{3/2} m)$, such that*

$$\operatorname{dist}(\mathbf{z}^t, \mathbf{z}^\star) \le C_0\gamma \left(1 - \frac{\eta}{16}\right)^{t-T_\gamma}, \qquad t \ge T_\gamma. \tag{7}$$

Moreover, for every $0 \le t \le T_\gamma$, $c\log^{-2} m \le \|\mathbf{h}^t\|_2 \|\mathbf{x}^t\|_2 \le C$.

Proof. The proof is provided in Appendices A–F. □

The theorem gives a precise interpretation of the two stages shown in Figure 1.

- During Stage I, the recovery error changes slowly while the iterate approaches the solution set. This transient lasts for at most $T_\gamma = \mathcal{O}(\eta^{-1}\log m)$ iterations and ends when the iterate enters an $\mathcal{O}(\gamma)$ neighborhood of the solution set. Thus, within the admissible range in Theorem 1, a larger step size shortens Stage I.
- During Stage II, the quotient distance contracts with coefficient $1 - \eta/16$ at every iteration. Reducing the error from $\mathcal{O}(\gamma)$ to ε takes $\mathcal{O}(\eta^{-1}\log(\gamma/\varepsilon))$ additional iterations. A larger admissible step size therefore also accelerates the local linear-convergence phase.
- Combining the two stages, the algorithm reaches accuracy ε within $\mathcal{O}(\eta^{-1}\{\log m + \log(\gamma/\varepsilon)\})$ iterations. The lower bound on $\|\mathbf{h}^t\|_2 \|\mathbf{x}^t\|_2$ keeps the denominators in the normalized update (4) controlled throughout Stage I.

Remark 2 (Sample complexity)**.** *The cubic dependence on K is sufficient but not optimal. The empirical transition in Figure 4(b) suggests that the actual dependence on K is closer to linear. Closing this gap requires a sharper finite-sample analysis of Stage I and is left for future work.*

4 Why Random Initialization Works

Theorem 1 describes when the algorithm enters the local convergence region. We next explain how a random iterate reaches that region by considering the evolution of the parts of the two vectors that point toward the ground truth and the energy that remains outside those directions.

To expose the Stage-I dynamics, we use the scaling ambiguity to choose equal-norm representatives. Equalizing the norms of the two vectors makes the normalization denominators in the two updates coincide. This allows the average and difference of the two truth correlations to evolve separately.

Specifically, define $\widetilde{\mathbf{h}}^t = \sqrt{\|\mathbf{x}^t\|_2 / \|\mathbf{h}^t\|_2}\,\mathbf{h}^t$ and $\widetilde{\mathbf{x}}^t = \sqrt{\|\mathbf{h}^t\|_2 / \|\mathbf{x}^t\|_2}\,\mathbf{x}^t$. This change of coordinates leaves the recovered matrix unchanged, i.e., $\widetilde{\mathbf{h}}^t\widetilde{\mathbf{x}}^{t\mathsf{H}} = \mathbf{h}^t\mathbf{x}^{t\mathsf{H}}$. We then decompose these representatives as

$$p_t = \mathbf{h}^{\star\mathsf{H}}\widetilde{\mathbf{h}}^t, \quad \widetilde{\mathbf{h}}^t = p_t\mathbf{h}^\star + \widetilde{\mathbf{h}}_\perp^t, \qquad q_t = \mathbf{x}^{\star\mathsf{H}}\widetilde{\mathbf{x}}^t, \quad \widetilde{\mathbf{x}}^t = q_t\mathbf{x}^\star + \widetilde{\mathbf{x}}_\perp^t. \tag{8}$$

Here p_t and q_t are the correlations and represent the energy of the iterates along the directions of the true vectors, and the perpendicular terms $\widetilde{\mathbf{h}}_\perp^t$ and $\widetilde{\mathbf{x}}_\perp^t$ contain the energy orthogonal to the true directions. To consider the iteration of the updates, define

$$s_t = \frac{p_t + q_t}{2}, \qquad d_t = \frac{p_t - q_t}{2}, \tag{9}$$

where the average s_t is the common signal component shared by the two vectors, while d_t measures the mismatch between their correlations. Their separate population dynamics are derived below.

Define the scale-invariant geometric mean of the two vector norms by $r_t := \sqrt{\|\mathbf{h}^t\|_2\,\|\mathbf{x}^t\|_2}$. By construction, $\left\|\widetilde{\mathbf{h}}^t\right\|_2 = \|\widetilde{\mathbf{x}}^t\|_2 = r_t$. We summarize the signal and the remaining energy by

$$S_t = |s_t|^2, \quad D_t = |d_t|^2, \quad E_t = \frac{\left\|\widetilde{\mathbf{h}}_\perp^t\right\|_2^2 + \|\widetilde{\mathbf{x}}_\perp^t\|_2^2}{2}, \quad W_t = D_t + E_t. \tag{10}$$

Here, S_t is the useful signal energy and W_t is the undesired energy that combines the correlation mismatch D_t and the orthogonal energy E_t. We have the identity

$$r_t^2 = S_t + W_t, \tag{11}$$

which follows from balance and orthogonality.

For random unit-norm initialization, concentration on the complex sphere gives $|p_0|, |q_0| = \mathcal{O}(K^{-1/2})$, while the orthogonal projections retain nearly all of the two vector norms, so $E_0 = \Theta(1)$. Hence $S_0/W_0 = \mathcal{O}(K^{-1})$. Since $\langle \mathbf{h}^0\mathbf{x}^{0\mathsf{H}}, \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\rangle_{\mathrm{F}} = \overline{p}_0 q_0$, the initialization is weakly aligned with the truth. The iterate therefore begins far from the local convergence region, and its recovery error decreases slowly during the early part of Stage I. As the iteration proceeds, the mismatch and orthogonal components decay, the common signal component grows, and the ratio S_t/W_t increases. Once the signal becomes dominant, the recovery error drops rapidly, as observed in Figure 1. The next subsection makes this mechanism precise through the population dynamics.

4.1 Stage I: growth of the signal component

To understand the first stage, we begin with the population loss $F(\mathbf{h}, \mathbf{x}) = \left\|\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2$, which is the infinite-sample version of (2). For intuition, we consider its continuous-time gradient dynamics. By suppressing the time arguments, the four components obey

$$\dot{s} = (-1 + r^{-2})s, \quad \dot{d} = -(1 + r^{-2})d, \qquad \dot{\widetilde{\mathbf{h}}}_\perp = -\widetilde{\mathbf{h}}_\perp, \quad \dot{\widetilde{\mathbf{x}}}_\perp = -\widetilde{\mathbf{x}}_\perp. \tag{12}$$

Consequently,

$$\frac{d}{dt}\log\frac{S}{W} = \frac{2}{r^2}\left(1 + \frac{|d|^2}{W}\right). \tag{13}$$

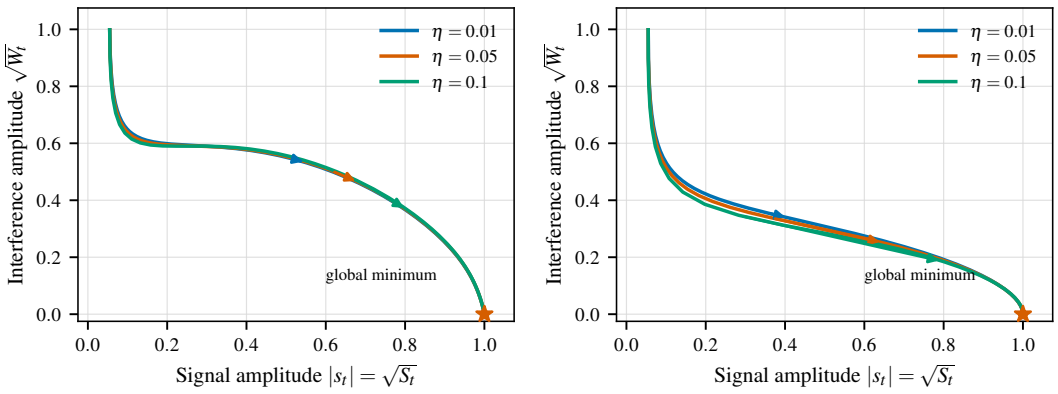
See Appendix A for the derivation of equations (12) and (13).

Equation (12) shows that the orthogonal components $\widetilde{\mathbf{h}}_\perp$ and $\widetilde{\mathbf{x}}_\perp$ decay at a fixed rate, and that the mismatch component d decays faster, with rate $1 + r^{-2}$. In contrast, the common signal component s is nearly stationary at initialization, when $r = 1$. As W decays, the identity $r^2 = S + W$ brings r below one, making the coefficient $-1 + r^{-2}$ positive and causing s to grow. Thus the signal remains weak while the undesired energy is being removed, which explains the slow initial decrease of the recovery error. Equation (13) shows that the ratio S/W increases strictly throughout it. The trajectory therefore moves from $S_0 \ll W_0$ to a state dominated by the signal component, and subsequently toward $S = 1, W = 0$.

With finitely many measurements, the empirical updates are perturbations of the population dynamics above. A uniform low-rank RIP bound controls the gradient perturbations over the entire Stage-I

trajectory. Under the sample-size condition of Theorem 1, these perturbations remain small, so the finite-sample state evolution follows the same signal-growth and interference-decay mechanism. The uniform approximation and the resulting discrete state evolution are proved in Appendices C–D.

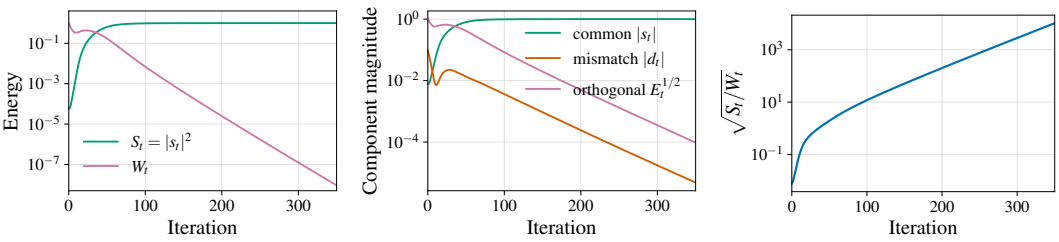
Figure 2(a) and Figure 2(b) compare the finite-sample and population trajectories in $(\sqrt{S_t}, \sqrt{W_t})$ for $K = 128$, $m = 10K$, and $\eta \in \{0.01, 0.05, 0.1\}$, using the same random initialization. Arrows indicate increasing iteration number. The point $(1, 0)$ represents the ground truth. It is observed that, starting from a random point, the trajectory first moves downward as the undesired energy decreases and then turns toward $(1, 0)$ as the signal component becomes dominant. Moreover, the finite-sample dynamics closely follow the population dynamics.



(a) Finite sample. (b) Population.

Figure 2: Finite-sample and population state trajectories.

To see how the individual components produce the trajectories in Figure 2, Figure 3 tracks their evolution for $K = 128$, $m = 10K$, and $\eta = 0.1$. Figure 3(a) shows that the useful energy S_t increases and the total undesired energy W_t decreases in a few iterations. Figure 3(b) further resolves the decay of W_t into the mismatch component $|d_t|$ and the orthogonal component $E_t^{1/2}$. In Figure 3(c), it is seen that the signal-to-interference ratio $\sqrt{S_t/W_t}$ increases through the iterations, which is consistent with the prediction in (13).



(a) Energy. (b) Components. (c) Signal-to-interference ratio.

Figure 3: Stage-I component dynamics.

4.2 Stage II: Local Contraction

At the end of Stage I, the updated vectors are close to the ground truth up to the scaling ambiguity. They are also sufficiently spread across the individual sensing vectors. These properties place the iterate in the local region studied by Ma et al. (2018). Within this region, each gradient step reduces the quotient distance (6) by a fixed proportion and keeps the next iterate in the same region. Repeating this local contraction gives the linear rate in (7). Appendix F verifies the required local conditions and gives the complete Stage-II argument.

5 Extension to Joint Blind Deconvolution and Demixing

BDD models settings in which several unknown signals pass through unknown channels and only their superposition is observed. Examples include multiuser wireless communication without explicit channel estimation (Ling & Strohmer, 2019) and simultaneous deconvolution and demixing in calcium imaging (Pnevmatikakis et al., 2016). Following the BDD formulation of Ling & Strohmer (2017), we extend the local mechanism above to recover multiple sources from a single set of measurements:

$$y_j = \sum_{i=1}^{n} \mathbf{b}_j^{\mathsf{H}} \mathbf{h}_i^{\star} \mathbf{x}_i^{\star \mathsf{H}} \mathbf{a}_{ij}, \qquad \mathbf{a}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(\mathbf{0}, \mathbf{I}_K), \quad 1 \le j \le m. \tag{14}$$

Without loss of generality, we assume that $\|\mathbf{h}_i^{\star}\|_2 = \|\mathbf{x}_i^{\star}\|_2 =: \sigma_i$, and define a condition number $\kappa := \max_i \sigma_i / \min_i \sigma_i$ to represent the spread of source strengths.

Let $\mathbf{z}_i := [\mathbf{h}_i^{\mathsf{H}}, \mathbf{x}_i^{\mathsf{H}}]^{\mathsf{H}}$ denote the state of source i. For the least-squares loss

$$f_{\mathrm{BDD}}(\{\mathbf{z}_i\}_{i=1}^n) := \sum_{j=1}^{m} \left| \sum_{i=1}^{n} \mathbf{b}_j^{\mathsf{H}} \mathbf{h}_i \mathbf{x}_i^{\mathsf{H}} \mathbf{a}_{ij} - y_j \right|^2, \tag{15}$$

we apply the BD update separately to each source:

$$\mathbf{h}_i^{t+1} = \mathbf{h}_i^t - \eta \frac{\nabla_{\mathbf{h}_i} f_{\mathrm{BDD}}}{\|\mathbf{x}_i^t\|_2^2}, \qquad \mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \eta \frac{\nabla_{\mathbf{x}_i} f_{\mathrm{BDD}}}{\|\mathbf{h}_i^t\|_2^2}, \qquad 1 \le i \le n. \tag{16}$$

Similar to the BD case, the denominators in (16) balance the updates of the two vectors within each source. Define $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$ and $\mathbf{Z}^{\star} = [\mathbf{z}_1^{\star}, \dots, \mathbf{z}_n^{\star}]$. The quotient distance of BDD is defined by

$$\mathrm{dist}_{\mathrm{BDD}}(\mathbf{Z}, \mathbf{Z}^{\star})^2 := \sum_{i=1}^{n} \mathrm{dist}(\mathbf{z}_i, \mathbf{z}_i^{\star})^2. \tag{17}$$

The following theorem states the local convergence guarantee for BDD. The precise definition of the local region and the technical conditions that keep the iterates inside it are given in Appendix G.

Theorem 3 (Local linear convergence of BDD)**.** *Let $\mathcal{R}$ be the good local region around the solution orbit specified in Appendix G. There exists a numerical constant $\eta_0 > 0$ such that, on the local event $\mathcal{E}_{\mathrm{loc}}$ defined there, the iterates of (16) initialized at $\mathbf{Z}^0 \in \mathcal{R}$ satisfy, for every $0 < \eta \le \eta_0$,*

$$\mathrm{dist}_{\mathrm{BDD}}(\mathbf{Z}^t, \mathbf{Z}^{\star}) \le \left(1 - \frac{\eta}{16}\right)^t \mathrm{dist}_{\mathrm{BDD}}(\mathbf{Z}^0, \mathbf{Z}^{\star}), \qquad t \ge 0. \tag{18}$$

The proof in Appendix G adapts the local-convergence argument of Ma et al. (2018) to the BDD setting. The convergence rate in (18) is independent of the condition number κ. This improves upon the κ-dependent convergence rate for the regularized gradient method of Ling & Strohmer (2019).

6 Numerical Experiments

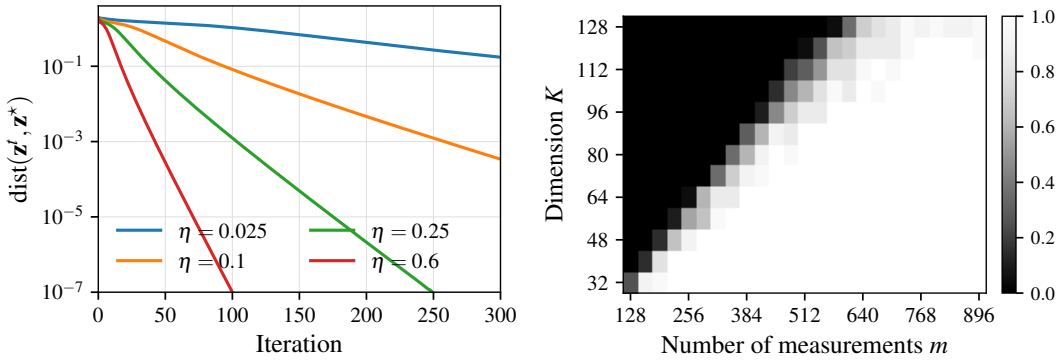
In this section, we present numerical experiments to illustrate the two-stage convergence of the randomly initialized Wirtinger flow for BD and the local convergence of BDD.

6.1 BD Experiments

Throughout this subsection, we sample the ground truth and the initial points uniformly from the complex unit sphere by normalizing independent standard complex Gaussian vectors. The equivalence follows from the unitary rotational invariance of the standard complex Gaussian distribution. We generate the Gaussian sensing vectors and the Fourier frame according to Section 2, and implement all experiments in Python 3.11.

We first study the effect of the step size. We fix $K = 128$ and $m = 10K$, and compare $\eta \in \{0.025, 0.1, 0.25, 0.6\}$ in Figure 4(a). It is observed that an increasing η shortens the initial transient and accelerates the subsequent linear-convergence phase. The two-stage shape agrees with the findings in Theorem 1.

We next study the impact of the dimension and the number of measurements on the recovery rate. We vary $K \in \{32, 40, \dots, 128\}$ and $m \in \{128, 160, \dots, 896\}$ independently, and declare a trial successful if $\left\|\mathbf{h}^t\mathbf{x}^{t\mathsf{H}} - \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \le 10^{-3}$ within 800 iterations. Each cell in Figure 4(b) averages 20 independent trials. The grayscale ranges from black for zero success probability to white for unit success probability. The transition from failure to recovery moves approximately linearly with K. This confirms the predicted improvement in recovery as the number of measurements grows, but it also indicates that the sufficient cubic scaling in Theorem 1 is conservative.



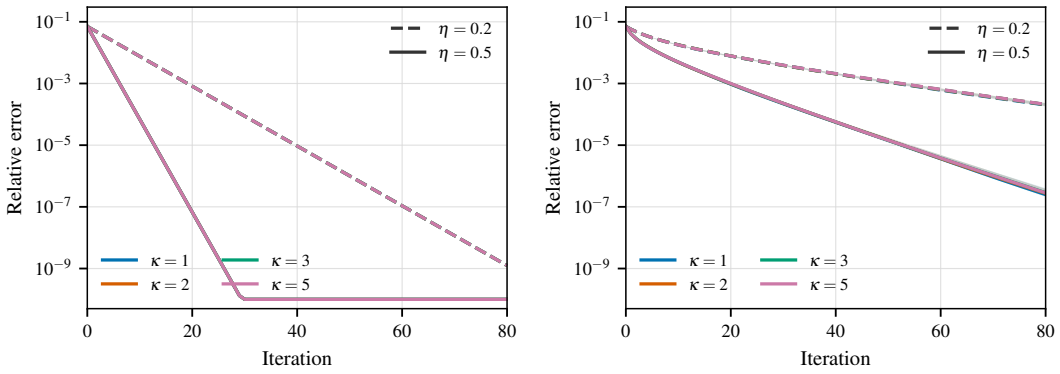
(a) Step-size stability.

(b) Empirical sample transition.

Figure 4: Step-size stability and empirical sample transition.

6.2 BDD CONVERGENCE

In this subsection, we test the local convergence rate of BDD in Section 5. To isolate the effect of the source condition number on this rate, we initialize every source inside the local region with the same relative perturbation. We set $n = 2$, $K = 50$, $m = 800$, $\sigma_1 = 1$, $\sigma_2 = \kappa$, and run 80 iterations for $\kappa \in \{1, 2, 3, 5\}$ and $\eta \in \{0.2, 0.5\}$. For every source i, we draw independent unit vectors $\boldsymbol{\varphi}_{h,i} \perp \mathbf{h}_i^\star$ and $\boldsymbol{\varphi}_{x,i} \perp \mathbf{x}_i^\star$, and form $\mathbf{h}_i^0 = \mathbf{h}_i^\star + 0.05\sigma_i\boldsymbol{\varphi}_{h,i}$ and $\mathbf{x}_i^0 = \mathbf{x}_i^\star + 0.05\sigma_i\boldsymbol{\varphi}_{x,i}$. Scaling the perturbation by σ_i gives every source the same relative initial error, while the orthogonal directions keep the two vectors balanced. Consequently, differences among the trajectories reflect their local contraction behavior rather than unequal starting errors. Following Ling & Strohmer (2019, equation (2.9)), we report the relative error $\mathrm{RE}_t := \sqrt{\sum_{i=1}^n \left\|\mathbf{h}_i^t\mathbf{x}_i^{t\mathsf{H}} - \mathbf{h}_i^\star\mathbf{x}_i^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2} / \sqrt{\sum_{i=1}^n \left\|\mathbf{h}_i^\star\mathbf{x}_i^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2}$.



(a) Population trajectories.

(b) Finite-sample trajectories.

Figure 5: BDD Stage-II convergence across source condition numbers.

We first use the population counterpart of (16) to isolate whether the local contraction rate depends on the source condition number. For every κ and both stepsizes, we start from the same matched local perturbations described above. Figure 5(a) shows that the four curves for different values of κ

coincide for each stepsize. It is observed that the algorithm with $\eta = 0.5$ converges faster than the one with $\eta = 0.2$. Thus the population convergence rate is insensitive to κ.

We next repeat the experiment with $m = 800$ to determine whether the same behavior persists in the finite-sample case. In Figure 5(b), each curve is the median of RE_t over the 20 matched trials, and its shaded band shows the 25th–75th percentile range at each iteration. Similar to the population case, the curves for $\eta = 0.5$ decay faster than those for $\eta = 0.2$, and the trajectories for different values of κ remain nearly overlapping at either stepsize. Therefore, the finite-sample convergence rate is also insensitive to the source condition number, which confirms our claim in Theorem 3.

Finally, we examine whether the same rate persists when the algorithm starts from a random point. We keep the finite-sample setting above and run 300 iterations. For each source, we sample $\mathbf{h}_i^0$ and $\mathbf{x}_i^0$ independently and uniformly from the complex unit sphere and then scale both vectors by σ_i. Figure 6 reports the median over 20 trials, with shaded 25th–75th percentile bands. For $\eta = 0.5$, the trajectories for all values of κ nearly overlap. For $\eta = 0.2$, the initial transient is more variable, especially at $\kappa = 5$, but the subsequent linear convergence part remain nearly parallel across κ. Thus, the linear convergence rate is insensitive to the source condition number even when the trajectory begins from a random initialization.

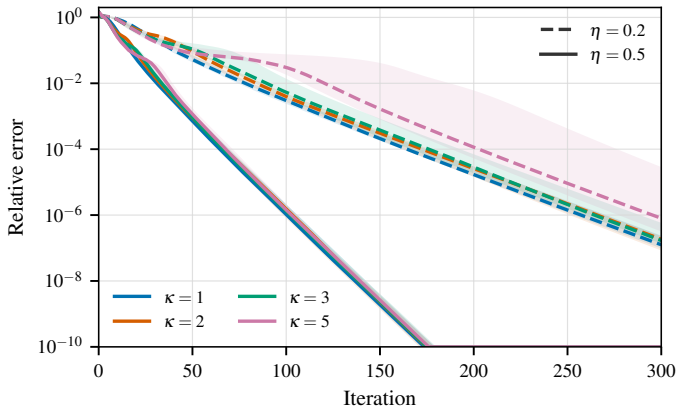


Figure 6: Finite-sample BDD convergence from random initialization.

7 Conclusion and Limitations

This paper establishes a global convergence guarantee for Wirtinger flow applied to BD from random initializations. The analysis explains the two stages observed in the iterates. During Stage I, the mismatch and orthogonal components decay while a weak common signal component becomes dominant. The iterate then enters the local region within $\mathcal{O}(\eta^{-1} \log m)$ iterations. During Stage II, the algorithm converges linearly to the ground truth. A low-rank RIP transfers the population dynamics to the finite-sample problem, and the numerical results illustrate the component evolution, step-size behavior, and sample transition. For the BDD extension, the same algorithm is applied sourcewise, yielding a local contraction coefficient that is independent of the source condition number. The population calculation and finite-sample experiments exhibit the same condition-number-insensitive behavior.

Several limitations remain. First, the sufficient sample size $m \gtrsim K^3 \log^{11} m$ is conservative. The approximately linear transition in Figure 4(b) suggests that a sharper analysis should exploit the particular Stage-I trajectory rather than control all low-rank directions uniformly. Establishing such a guarantee requires tighter finite-sample control of the common signal, mismatch, and the norms of the two vectors. Second, we only consider the noiseless case in a Fourier–Gaussian sensing model. Extending the result to noisy observations and more general structured frames might be possible. Third, our BDD guarantee is local. The experiment in Figure 6 suggests that the condition-number-insensitive linear rate also emerges from random initialization. Characterizing the preceding BDD trajectory dynamics and proving a corresponding global convergence guarantee are left for future work.

Reproducibility Statement

To ensure the reproducibility of our work, we have provided a detailed description of the experimental setup in Section 6. Specifically, all model hyperparameters, initialization strategies, and implementation details are explicitly provided. The programming language and its exact version are also clearly stated. We believe these details are sufficient to reproduce our results.

AI Use Statement

During the preparation of this manuscript, AI-assisted tools were used for language editing and stylistic refinement. They were also utilized to assist with LaTeX formatting, including the typesetting of equations, figure placements, and citations. The authors carefully reviewed all AI-generated suggestions and take full responsibility for the final content of this paper.

References

Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014. doi: 10.1109/TIT.2013.2294644.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3880–3888. Curran Associates, Inc., 2016.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1–2): 5–37, 2019. doi: 10.1007/s10107-019-01363-6.

Jialin Dong and Yuanming Shi. Nonconvex demixing from bilinear measurements. *IEEE Transactions on Signal Processing*, 66(19):5152–5166, 2018. doi: 10.1109/TSP.2018.2864660.

Jialin Dong and Yuanming Shi. Blind demixing via Wirtinger flow with random initialization. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 362–370. PMLR, 2019.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1233–1242. PMLR, 2017.

Dimitris Hatzinakos. Blind deconvolution channel identification and equalization. In *Control and Dynamic Systems*, volume 68, pp. 279–331. Elsevier, 1995. doi: 10.1016/S0090-5267(06)80043-5.

Wen Huang and Paul Hand. Blind deconvolution by a steepest descent algorithm on a quotient manifold. *SIAM Journal on Imaging Sciences*, 11(4):2757–2785, 2018. doi: 10.1137/17M1151390.

Deepa Kundur and Dimitris Hatzinakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, 1996. doi: 10.1109/79.489268.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 47(3): 893–934, 2019. doi: 10.1016/j.acha.2018.01.001.

Yanjun Li and Yoram Bresler. Multichannel sparse blind deconvolution on the sphere. *IEEE Transactions on Information Theory*, 65(11):7415–7436, 2019. doi: 10.1109/TIT.2019.2928576.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47. PMLR, 2018.

Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel Lee. Blind channel identification for speech dereverberation using ℓ_1-norm sparse learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017. doi: 10.1109/TIT.2017.2701342.

Shuyang Ling and Thomas Strohmer. Regularized gradient descent: A non-convex recipe for fast joint blind deconvolution and demixing. *Information and Inference: A Journal of the IMA*, 8(1): 1–49, 2019. doi: 10.1093/imaiai/iax022.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3345–3354. PMLR, 2018.

Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021. doi: 10.1109/TSP.2021.3051425.

Eftychios A. Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A. Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha Ahrens, Randy Bruno, Thomas M. Jessell, Darcy S. Peterka, Rafael Yuste, and Liam Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016. doi: 10.1016/j.neuron.2015.11.037.

Qing Qu, Xiao Li, and Zhihui Zhu. Exact recovery of multichannel sparse blind deconvolution via gradient descent. *SIAM Journal on Imaging Sciences*, 13(3):1630–1652, 2020. doi: 10.1137/19M1291327.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23831–23843. Curran Associates, Inc., 2021.

Tian Ye and Simon Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1429–1439. Curran Associates, Inc., 2021.

Yuqian Zhang, Han-Wen Kuo, and John Wright. Structured local optima in sparse blind deconvolution. *IEEE Transactions on Information Theory*, 66(1):419–452, 2020. doi: 10.1109/TIT.2019.2940657.

Yuqian Zhang, Yenson Lau, Han-Wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):999–1008, 2021. doi: 10.1109/TPAMI.2019.2939237.

A Population coordinates

This section derives the population dynamics in the equal-norm coordinates defined in Section 4.

For each $c \in \mathbb{C} \setminus \{0\}$, define the scaling map $\mathcal{T}_c : \mathbb{C}^K \times \mathbb{C}^K \to \mathbb{C}^K \times \mathbb{C}^K$ by $\mathcal{T}_c(\mathbf{h}, \mathbf{x}) = (\mathbf{h}/\overline{c}, c\mathbf{x})$. Let $\mathcal{G} : \mathbb{C}^K \times \mathbb{C}^K \to \mathbb{C}^K \times \mathbb{C}^K$ be the one-step update map induced by (4), so that, with the notation from Section 2, $\mathbf{z}^{t+1} = \mathcal{G}(\mathbf{z}^t)$. The residual l_j is invariant under $\mathcal{T}_c$, while the gradients obey

$$\nabla_{\mathbf{h}} f(\mathcal{T}_c(\mathbf{h}, \mathbf{x})) = c \nabla_{\mathbf{h}} f(\mathbf{h}, \mathbf{x}), \qquad \nabla_{\mathbf{x}} f(\mathcal{T}_c(\mathbf{h}, \mathbf{x})) = \nabla_{\mathbf{x}} f(\mathbf{h}, \mathbf{x})/\overline{c}. \tag{19}$$

Combining these identities with $\|c\mathbf{x}\|_2^2 = |c|^2 \|\mathbf{x}\|_2^2$ and $\|\mathbf{h}/\overline{c}\|_2^2 = |c|^{-2} \|\mathbf{h}\|_2^2$ gives

$$\mathcal{G}(\mathcal{T}_c(\mathbf{h}, \mathbf{x})) = \mathcal{T}_c(\mathcal{G}(\mathbf{h}, \mathbf{x})). \tag{20}$$

Thus the update preserves scaling equivalence. Selecting the equal-norm representative $(\widetilde{\mathbf{h}}^t, \widetilde{\mathbf{x}}^t)$ after each raw update therefore preserves the entire sequence of recovered rank-one matrices. To keep

the formulas readable, throughout the analysis of Stage-I we write the equal-norm representative as $(\mathbf{h}^t, \mathbf{x}^t)$ with a little abuse of notation.

We first give the Fourier-frame identities that will be used repeatedly: $\sum_{j=1}^m \mathbf{b}_j \mathbf{b}_j^{\mathsf{H}} = \mathbf{I}_K$ and $\|\mathbf{b}_j\|_2^2 = K/m$. For the calculations below, define the sensing map $\mathcal{A} : \mathbb{C}^{K\times K} \to \mathbb{C}^m$ by $[\mathcal{A}(\mathbf{Z})]_j = \mathbf{b}_j^{\mathsf{H}} \mathbf{Z} \mathbf{a}_j$ for $1 \le j \le m$.

The population loss and the corresponding Wirtinger gradients are

$$F(\mathbf{h}, \mathbf{x}) := \mathbb{E} f(\mathbf{h}, \mathbf{x}) = \left\| \mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star \right\|_{\mathrm{F}}^2, \tag{21}$$

$$\nabla_{\mathbf{h}} F(\mathbf{h}, \mathbf{x}) = \|\mathbf{x}\|_2^2 \, \mathbf{h} - \left(\mathbf{x}^{\star\mathsf{H}} \mathbf{x}\right) \mathbf{h}^\star, \tag{22}$$

$$\nabla_{\mathbf{x}} F(\mathbf{h}, \mathbf{x}) = \|\mathbf{h}\|_2^2 \, \mathbf{x} - \left(\mathbf{h}^{\star\mathsf{H}} \mathbf{h}\right) \mathbf{x}^\star. \tag{23}$$

For an equal-norm pair, decompose each factor vector into its component along the truth and its orthogonal component:

$$\begin{aligned} p &= \mathbf{h}^{\star\mathsf{H}} \mathbf{h}, & \mathbf{h} &= p\mathbf{h}^\star + \mathbf{h}_\perp, & \mathbf{h}^{\star\mathsf{H}} \mathbf{h}_\perp &= 0, \\ q &= \mathbf{x}^{\star\mathsf{H}} \mathbf{x}, & \mathbf{x} &= q\mathbf{x}^\star + \mathbf{x}_\perp, & \mathbf{x}^{\star\mathsf{H}} \mathbf{x}_\perp &= 0 \end{aligned} \tag{24}$$

and the common signal and mismatch components are

$$s = \frac{p+q}{2}, \qquad d = \frac{p-q}{2}. \tag{25}$$

We use the corresponding energy variables

$$S = |s|^2, \quad D = |d|^2, \quad U = \|\mathbf{h}_\perp\|_2^2, \quad V = \|\mathbf{x}_\perp\|_2^2, \quad W = D + \frac{U+V}{2}. \tag{26}$$

Balance implies the identity

$$r^2 = S + W. \tag{27}$$

The next lemma records the resulting population flow. It is the basic calculation behind the two-stage interpretation in Section 4.

Lemma 4. *Consider the normalized population gradient flow*

$$\dot{\mathbf{h}} = -\frac{\nabla_{\mathbf{h}} F(\mathbf{h}, \mathbf{x})}{\|\mathbf{x}\|_2^2}, \qquad \dot{\mathbf{x}} = -\frac{\nabla_{\mathbf{x}} F(\mathbf{h}, \mathbf{x})}{\|\mathbf{h}\|_2^2}. \tag{28}$$

If the two factor vectors initially have equal norms, this equality is preserved and

$$\dot{s} = (-1 + r^{-2}) s, \qquad \dot{d} = (-1 - r^{-2}) d, \tag{29}$$

$$\dot{\mathbf{h}}_\perp = -\mathbf{h}_\perp, \qquad \dot{\mathbf{x}}_\perp = -\mathbf{x}_\perp. \tag{30}$$

Consequently,

$$\dot{S} = 2(-1 + r^{-2}) S, \tag{31}$$

$$\dot{W} = -2W - \frac{2D}{r^2}, \tag{32}$$

$$\frac{d}{dt} \log \frac{S}{W} = \frac{2}{r^2} \left(1 + \frac{D}{W}\right). \tag{33}$$

Proof. Substitution of (22) and (23) into (28) gives

$$\dot{\mathbf{h}} = -\mathbf{h} + \frac{q}{\|\mathbf{x}\|_2^2} \mathbf{h}^\star, \qquad \dot{\mathbf{x}} = -\mathbf{x} + \frac{p}{\|\mathbf{h}\|_2^2} \mathbf{x}^\star.$$

When $\|\mathbf{h}\|_2 = \|\mathbf{x}\|_2 = r$,

$$\frac{d}{dt} \|\mathbf{h}\|_2^2 = -2r^2 + \frac{2\,\mathrm{Re}(\overline{p} q)}{r^2} = \frac{d}{dt} \|\mathbf{x}\|_2^2,$$

so the balance condition is preserved by the flow. Projecting onto the truth directions and their orthogonal complements gives (29)–(30). Differentiating the definitions of S and W gives (31) and (32); subtracting their logarithmic derivatives gives (33). □

B Population Stage I for blind deconvolution

This section turns the population flow into quantitative bounds in Stage-I. The useful energy is initially much smaller than the undesired energy, so we first control their ratio and show that their common norm r stays away from zero. After the two energies become comparable, the useful component grows to constant size and the trajectory enters the region of local convergence. The following lemma summarizes these three facts.

Lemma 5. *Suppose $r(0) = 1$, $S_0 > 0$, $E_0 := (U_0 + V_0)/2 > 0$, and $z_0 := S_0/W_0 < 1$. Put $L_0 = 1 - \log z_0 - z_0$. If $T_\times = \inf\{t : S(t) \geq W(t)\}$, then*

$$\frac{1}{2}\log\left(1 + \frac{E_0 L_0}{2}\right) \leq T_\times \leq \frac{1}{2}\log(1 + W_0 L_0), \tag{34}$$

and

$$\frac{E_0}{1 + W_0 L_0} \leq \inf_{t \geq 0} r(t)^2 \leq \frac{2W_0}{1 + (E_0/2)L_0}. \tag{35}$$

For every $0 < \varepsilon < c$, the flow reaches

$$W(t) \leq C\varepsilon^2, \qquad |S(t) - 1| \leq C\varepsilon \tag{36}$$

by time $T_\times + C\log(1/\varepsilon)$.

Proof. Let $z = S/W$. The equation for W immediately gives the two-sided estimate

$$E_0 e^{-2t} \leq W(t) \leq W_0 e^{-2t}. \tag{37}$$

Since $r^2 = (1 + z)W$, the ratio equation (33) yields

$$\frac{d}{dt}(\log z + z) = (1 + z)\frac{d}{dt}\log z \geq \frac{2}{W} \geq \frac{2e^{2t}}{W_0}, \tag{38}$$

$$\frac{d}{dt}(\log z + z) \leq \frac{4}{W} \leq \frac{4e^{2t}}{E_0}. \tag{39}$$

Integrating these inequalities until z reaches one proves (34).

We next bound the common vector norm. From the population equations,

$$\frac{d}{dt}r^2 = -2r^2 + \frac{2(S - D)}{r^2} \leq 2(1 - r^2).$$

Since $r(0) = 1$, comparison gives $r(t)^2 \leq 1$. Before $T_\times$, (37) and the upper bound in (34) imply

$$r(t)^2 \geq W(t) \geq E_0 e^{-2T_\times} \geq \frac{E_0}{1 + W_0 L_0}.$$

At $T_\times$, $S = W$, and S is nondecreasing because $r^2 \leq 1$. The same lower bound therefore holds after $T_\times$. Conversely,

$$r(T_\times)^2 = 2W(T_\times) \leq 2W_0 e^{-2T_\times} \leq \frac{2W_0}{1 + (E_0/2)L_0},$$

which proves the upper bound in (35). The lower bound was obtained above.

It remains to show that the flow enters the local convergence region. After $T_\times$, (33) gives $S/W \geq e^{2(t - T_\times)}$, while $W(t) \leq e^{-2(t - T_\times)}$. If $S \leq 1/4$, then $W \leq S$ and

$$\dot{S} = 2S\frac{1 - S - W}{S + W} \geq \frac{1}{2},$$

so S reaches $1/4$ in constant time. Thereafter, for $e = 1 - S$,

$$\dot{e} = -\frac{2S}{S + W}(e - W) \leq -ce + CW.$$

If $t_0 \leq T_\times + C$ is the first time $S(t_0) = 1/4$, variation of constants and $W(t) \leq e^{-2(t - T_\times)}$ give

$$e(t) \leq e^{-c(t - t_0)}e(t_0) + C\int_{t_0}^{t} e^{-c(t - \tau)}W(\tau)\,d\tau \leq Ce^{-c(t - t_0)} + Ce^{-c(t - T_\times)}.$$

Thus both $W(t) \leq C\varepsilon^2$ and $e(t) \leq C\varepsilon$ hold once $t - T_\times \geq C\log(1/\varepsilon)$. This proves (36). □

C UNIFORM CONTROL OF THE EMPIRICAL GRADIENTS

The population calculations in the above sections describe the idealized infinite-sample update. To compare them with the empirical algorithm, we need uniform control of the sensing map over all low-rank directions. We first state the rank-restricted isometry property and prove it for the Fourier–Gaussian sensing map. We then use this property to bound the difference between the empirical and population gradients.

Definition 1 (Rank-r restricted isometry property)**.** *For $r \geq 1$, the sensing map $\mathcal{A}$ is said to have rank-r restricted isometry constant δ_r if*

$$\left| \|\mathcal{A}(\mathbf{Z})\|_2^2 - \|\mathbf{Z}\|_{\mathrm{F}}^2 \right| \leq \delta_r \|\mathbf{Z}\|_{\mathrm{F}}^2 \quad \textit{for every } \mathbf{Z} \textit{ with } \operatorname{rank}(\mathbf{Z}) \leq r. \tag{40}$$

The next lemma provides the uniform rank-four RIP used to compare the empirical and population gradients.

Lemma 6. *Let $0 < \delta < 1/2$. Under the assumptions of Section 2, if $m \geq C\delta^{-2}K^2 \log(CK/\delta)$, then $\mathcal{A}$ has rank-4 RIP constant at most δ with probability at least $1 - 2\exp(-c\delta^2 m/K)$.*

Proof. Fix a matrix $\mathbf{Z}$ with $\|\mathbf{Z}\|_{\mathrm{F}} = 1$ and put $\mathbf{v}_j = \mathbf{Z}^{\mathsf{H}}\mathbf{b}_j$. Then

$$\|\mathcal{A}(\mathbf{Z})\|_2^2 - \|\mathbf{Z}\|_{\mathrm{F}}^2 = \sum_{j=1}^{m} \left(|\mathbf{a}_j^{\mathsf{H}}\mathbf{v}_j|^2 - \|\mathbf{v}_j\|_2^2 \right),$$

where we used the Fourier-frame identity $\sum_j \left\|\mathbf{Z}^{\mathsf{H}}\mathbf{b}_j\right\|_2^2 = \|\mathbf{Z}\|_{\mathrm{F}}^2$. For each summand, complex Gaussian concentration gives

$$\left\| |\mathbf{a}_j^{\mathsf{H}}\mathbf{v}_j|^2 - \|\mathbf{v}_j\|_2^2 \right\|_{\psi_1} \leq C \|\mathbf{v}_j\|_2^2 .$$

The deterministic Fourier norms also imply

$$\max_j \|\mathbf{v}_j\|_2^2 \leq \frac{K}{m}, \qquad \sum_j \|\mathbf{v}_j\|_2^4 \leq \frac{K}{m} \sum_j \|\mathbf{v}_j\|_2^2 = \frac{K}{m}.$$

Bernstein's inequality therefore gives the following bound for a fixed matrix:

$$\mathbb{P}\left(\left| \|\mathcal{A}(\mathbf{Z})\|_2^2 - 1 \right| > u \right) \leq 2\exp(-cu^2 m/K), \qquad 0 < u < 1. \tag{41}$$

We now extend the bound to all rank-four matrices. The unit sphere of rank-4 complex $K \times K$ matrices admits an ϵ-net in Frobenius norm with cardinality at most $(C/\epsilon)^{16K}$. This follows, for example, by writing a rank-4 matrix as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$, covering the two complex Stiefel factors and the unit sphere of singular values separately, and multiplying the three covering numbers. Let $\mathcal{N}_\epsilon$ denote such an ϵ-net, and apply (41) with $u = \delta/4$. A union bound gives

$$\max_{\mathbf{Z}_0 \in \mathcal{N}_\epsilon} \left| \|\mathcal{A}(\mathbf{Z}_0)\|_2^2 - 1 \right| \leq \delta/4 \tag{42}$$

except on an event of probability at most $2\exp\{16K\log(C/\epsilon) - c\delta^2 m/K\}$.

It remains to extend (42) from the finite net to every rank-four matrix of unit Frobenius norm. Define

$$M_4 := \sup_{\substack{\operatorname{rank}(\mathbf{Z}) \leq 4 \\ \|\mathbf{Z}\|_{\mathrm{F}} = 1}} \|\mathcal{A}(\mathbf{Z})\|_2 , \qquad M_8 := \sup_{\substack{\operatorname{rank}(\mathbf{Z}) \leq 8 \\ \|\mathbf{Z}\|_{\mathrm{F}} = 1}} \|\mathcal{A}(\mathbf{Z})\|_2 .$$

Splitting the singular-value decomposition of a rank-8 matrix into two rank-4 blocks shows $M_8 \leq \sqrt{2}M_4$. Given a rank-4 unit matrix $\mathbf{Z}$, choose $\mathbf{Z}_0 \in \mathcal{N}_\epsilon$ with $\|\mathbf{Z} - \mathbf{Z}_0\|_{\mathrm{F}} \leq \epsilon$. Since $\mathbf{Z} - \mathbf{Z}_0$ has rank at most 8,

$$\|\mathcal{A}(\mathbf{Z})\|_2 \leq \sqrt{1 + \delta/4} + M_8\epsilon \leq \sqrt{1 + \delta/4} + \sqrt{2}\epsilon M_4.$$

Taking the supremum and choosing $\epsilon = c_1\delta$ with c_1 sufficiently small yields $M_4 \leq 1 + C\delta$. The reverse triangle inequality gives, in the same way,

$$\|\mathcal{A}(\mathbf{Z})\|_2 \geq \sqrt{1 - \delta/4} - \sqrt{2}\epsilon M_4 \geq 1 - C\delta.$$

Reducing c_1 once more and rescaling the numerical value of δ proves (40). Finally, the assumed lower bound on m makes the net failure probability no larger than $2\exp(-c\delta^2 m/K)$. □

The next lemma provides a uniform approximation of the two empirical gradients by their population counterparts.

Lemma 7. *On the event of Lemma 6, for every* $\mathbf{h}, \mathbf{x} \in \mathbb{C}^K$,

$$\|\nabla_{\mathbf{h}} f(\mathbf{h}, \mathbf{x}) - \nabla_{\mathbf{h}} F(\mathbf{h}, \mathbf{x})\|_2 \le 2\delta \left\|\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star\right\|_{\mathrm{F}} \|\mathbf{x}\|_2 , \tag{43}$$

$$\|\nabla_{\mathbf{x}} f(\mathbf{h}, \mathbf{x}) - \nabla_{\mathbf{x}} F(\mathbf{h}, \mathbf{x})\|_2 \le 2\delta \left\|\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star\right\|_{\mathrm{F}} \|\mathbf{h}\|_2 . \tag{44}$$

Proof. By the complex polarization identity, the rank-4 RIP implies

$$|\langle \mathcal{A}(\mathbf{Z}), \mathcal{A}(\mathbf{Y})\rangle - \langle \mathbf{Z}, \mathbf{Y}\rangle_{\mathrm{F}}| \le 2\delta \|\mathbf{Z}\|_{\mathrm{F}} \|\mathbf{Y}\|_{\mathrm{F}} \tag{45}$$

whenever $\mathrm{rank}(\mathbf{Z}) + \mathrm{rank}(\mathbf{Y}) \le 4$, and applying this estimate with $\mathbf{Z} = \mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star$, $\mathbf{Y} = \mathbf{u}\mathbf{x}^{\mathsf{H}}$ or $\mathbf{Y} = \mathbf{h}\mathbf{u}^{\mathsf{H}}$, followed by a supremum over unit vectors $\mathbf{u}$, gives both displayed gradient bounds. □

We next record a standard concentration fact for complex Gaussian vectors. It will be used later to verify the incoherence and initialization conditions in the local-convergence argument. A Gaussian norm tail bound followed by a union bound over the m sensing vectors gives, with probability at least $1 - Cme^{-cK}$,

$$\max_{1\le j\le m} \|\mathbf{a}_j\|_2 \le 3\sqrt{K}. \tag{46}$$

D FINITE-SAMPLE STAGE-I DYNAMICS

The preceding gradient bound in Lemma 7 allows us to compare one empirical update with one population update. We now express this comparison in the four Stage-I coordinates. The terms ξ_s, ξ_d contain the finite-sample error and $\boldsymbol{\xi}_{\mathbf{h}}, \boldsymbol{\xi}_{\mathbf{x}}$ denote the second-order change caused by selecting the equal-norm representative after an update. The next lemma gives the one-step recursions in Stage I.

Lemma 8. *Suppose an equal-norm representative satisfies* $\ell^{-1} \le r^2 \le 4$, *where* $\ell \ge 2$, *and let* $\eta \le c\ell^{-2}$ *and* $\delta \le c\ell^{-2}$. *Suppose also that the conclusions of Lemma 7 hold. Apply one empirical update and then select the equal-norm representative of the updated pair of factor vectors. The coordinates in (25) satisfy*

$$s^+ = (1 - \eta + \eta r^{-2})s + \xi_s, \qquad d^+ = (1 - \eta - \eta r^{-2})d + \xi_d, \tag{47}$$

$$\mathbf{h}_\perp^+ = (1-\eta)\mathbf{h}_\perp + \boldsymbol{\xi}_{\mathbf{h}}, \qquad \mathbf{x}_\perp^+ = (1-\eta)\mathbf{x}_\perp + \boldsymbol{\xi}_{\mathbf{x}}, \tag{48}$$

where

$$|\xi_s| \le C\frac{\eta\delta}{r} + C\frac{\eta^2}{r^4}|s|, \tag{49}$$

$$|\xi_d| + \|\boldsymbol{\xi}_{\mathbf{h}}\|_2 + \|\boldsymbol{\xi}_{\mathbf{x}}\|_2 \le C\frac{\eta\delta}{r} + C\frac{\eta^2}{r^4}\left(|d| + \|\mathbf{h}_\perp\|_2 + \|\mathbf{x}_\perp\|_2\right). \tag{50}$$

The common norm after this selection satisfies

$$\left| r_+^2 - \left[(1-\eta)^2 r^2 + 2\eta(1-\eta)\frac{\mathrm{Re}(\overline{p}q)}{r^2} + \eta^2\frac{|p|^2+|q|^2}{2r^4}\right]\right| \le C\eta\delta + C\eta^2\delta^2/r^2. \tag{51}$$

Proof. We first analyze the raw update before selecting its equal-norm representative. (22)–(23) and Lemma 7 give

$$\widehat{\mathbf{h}}^+ = (1-\eta)\mathbf{h} + \frac{\eta q}{r^2}\mathbf{h}^\star + \mathbf{e}_{\mathbf{h}}, \qquad \|\mathbf{e}_{\mathbf{h}}\|_2 \le C\eta\delta/r, \tag{52}$$

$$\widehat{\mathbf{x}}^+ = (1-\eta)\mathbf{x} + \frac{\eta p}{r^2}\mathbf{x}^\star + \mathbf{e}_{\mathbf{x}}, \qquad \|\mathbf{e}_{\mathbf{x}}\|_2 \le C\eta\delta/r. \tag{53}$$

The residual is bounded by $\left\|\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star\right\|_{\mathrm{F}} \le r^2 + 1 \le 5$. Write $e_p = \mathbf{h}^{\star\mathsf{H}}\mathbf{e}_{\mathbf{h}}$ and $e_q = \mathbf{x}^{\star\mathsf{H}}\mathbf{e}_{\mathbf{x}}$. Let $\mathbf{P}_{\mathbf{h}} = \mathbf{I}_K - \mathbf{h}^\star\mathbf{h}^{\star\mathsf{H}}$ and $\mathbf{P}_{\mathbf{x}} = \mathbf{I}_K - \mathbf{x}^\star\mathbf{x}^{\star\mathsf{H}}$. Projecting the raw update onto the truth directions and their orthogonal complements gives (52)–(53) gives

$$\widehat{p} = (1-\eta)p + \eta q/r^2 + e_p, \qquad \widehat{q} = (1-\eta)q + \eta p/r^2 + e_q,$$

$$\widehat{\mathbf{h}}_\perp = (1-\eta)\mathbf{h}_\perp + \mathbf{P}_{\mathbf{h}}\mathbf{e}_{\mathbf{h}}, \qquad \widehat{\mathbf{x}}_\perp = (1-\eta)\mathbf{x}_\perp + \mathbf{P}_{\mathbf{x}}\mathbf{e}_{\mathbf{x}}.$$

Consequently,

$$\widehat{s} = (1 - \eta + \eta/r^2)s + (e_p + e_q)/2, \tag{54}$$

$$\widehat{d} = (1 - \eta - \eta/r^2)d + (e_p - e_q)/2. \tag{55}$$

The raw update can produce two factor vectors with unequal norms. To quantify this difference, we expand their squared norms and collect every term involving an empirical gradient error into ρ:

$$\left\|\widehat{\mathbf{h}}^+\right\|_2^2 - \left\|\widehat{\mathbf{x}}^+\right\|_2^2 = \frac{\eta^2}{r^4}(|q|^2 - |p|^2) + \rho, \qquad |\rho| \le C\eta\delta + C\eta^2\delta^2/r^2. \tag{56}$$

Since $|q|^2 - |p|^2 = -4\,\mathrm{Re}(s\overline{d})$, the term that remains after omitting the empirical gradient errors is bounded by a constant times $|s||d|$. Under the assumptions of the lemma, both updated norms lie between $r/2$ and $2r$: indeed, the displacement in each line of (52)–(53) is at most $\eta r + \eta/r + C\eta\delta/r \le r/2$ after reducing the numerical constant. For positive $A, B \in [r/2, 2r]$,

$$\left|\sqrt{A/B} - 1\right| = \frac{|A - B|}{\sqrt{B}(\sqrt{A} + \sqrt{B})} \le C\frac{|A^2 - B^2|}{r^2}.$$

Apply this with $A = \left\|\widehat{\mathbf{h}}^+\right\|_2$ and $B = \|\widehat{\mathbf{x}}^+\|_2$ to obtain

$$c_+ = \left(\left\|\widehat{\mathbf{h}}^+\right\|_2 / \left\|\widehat{\mathbf{x}}^+\right\|_2\right)^{1/2}, \qquad |c_+ - 1| \le C\frac{\eta^2|s||d|}{r^6} + C\frac{\eta\delta}{r^2}. \tag{57}$$

Here we used (56) and $|\rho|/r^2 \le C\eta\delta/r^2 + C\eta^2\delta^2/r^4$; the last term is absorbed into the first under the assumed bounds on r, η, and δ.

It remains to select the equal-norm representative of the updated pair. By the definition of c_+, $p^+ = \widehat{p}/c_+$ and $q^+ = c_+\widehat{q}$,

$$s^+ = \frac{c_+^{-1} + c_+}{2}\widehat{s} + \frac{c_+^{-1} - c_+}{2}\widehat{d}, \qquad d^+ = \frac{c_+^{-1} - c_+}{2}\widehat{s} + \frac{c_+^{-1} + c_+}{2}\widehat{d}.$$

Since $|c_+^{-1} + c_+ - 2| \le C|c_+ - 1|^2$ and $|c_+^{-1} - c_+| \le C|c_+ - 1|$, substitution of (54)–(57), together with $|s|, |d|, \|\mathbf{h}_\perp\|_2, \|\mathbf{x}_\perp\|_2 \le r$, gives (49)–(50). The perpendicular components follow in the same way from $\mathbf{h}_\perp^+ = \widehat{\mathbf{h}}_\perp/c_+$ and $\mathbf{x}_\perp^+ = c_+\widehat{\mathbf{x}}_\perp$.

Finally, put $R_h = \left\|\widehat{\mathbf{h}}^+\right\|_2^2$ and $R_x = \|\widehat{\mathbf{x}}^+\|_2^2$. Since $r_+^2 = \sqrt{R_h R_x}$ and

$$0 \le \frac{R_h + R_x}{2} - \sqrt{R_h R_x} = \frac{(R_h - R_x)^2}{2(\sqrt{R_h} + \sqrt{R_x})^2},$$

the difference between r_+^2 and $(R_h + R_x)/2$ is bounded using (56). Expanding $(R_h + R_x)/2$ directly from (52)–(53) yields the three population terms in (51); all terms containing $\mathbf{e_h}, \mathbf{e_x}$ are at most $C\eta\delta + C\eta^2\delta^2/r^2$. This proves (51). □

The next lemma shows that the local coordinate error contracts after each empirical-gradient step.

Lemma 9. *There are numerical constants $c_0, c_1, c_2 > 0$ such that the following holds on the event of Lemma 7. Suppose an equal-norm pair satisfies $\mathcal{E} := (|s| - 1)^2 + W \le c_0^2$, and let $0 < \eta \le c_1$ and $0 < \delta \le c_1$. Then one empirical-gradient step, followed by selecting the equal-norm representative, satisfies*

$$\mathcal{E}^+ \le (1 - c_2\eta)\mathcal{E}. \tag{58}$$

Moreover,

$$\mathrm{dist}(\mathbf{z}, \mathbf{z}^\star)^2 \le C\mathcal{E}. \tag{59}$$

Proof. Near the solution, the residual is proportional to the coordinate error. Keeping this dependence in the finite-sample gradient bound removes the additive term that would otherwise prevent the estimate from converging to zero. We use this sharper bound to prove contraction of the local error $\mathcal{E}$. By equivariance, multiply both factor vectors by a common unit-modulus phase so that $s \ge 0$ is

real. This phase rotation preserves W, the recovered matrix, and the distance. Write $e = s - 1$. The assumed local-energy bound and $r^2 = s^2 + W$ imply

$$1/2 \le r^2 \le 2, \qquad |e| \le c_0, \qquad W \le c_0^2. \tag{60}$$

Expanding $\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star$ using (24) and $p = s + d$, $q = s - d$, and then applying the triangle inequality gives

$$\left\|\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{X}^\star\right\|_{\mathrm{F}} \le C\big(|e| + |d| + \|\mathbf{h}_\perp\|_2 + \|\mathbf{x}_\perp\|_2\big) \le C\sqrt{\mathcal{E}}. \tag{61}$$

Consequently, the two empirical-gradient errors in (52)–(53) now obey the sharper bound

$$\|\mathbf{e}_\mathbf{h}\|_2 + \|\mathbf{e}_\mathbf{x}\|_2 \le C\eta\delta\sqrt{\mathcal{E}}. \tag{62}$$

We first repeat the coordinate calculation using this sharper residual bound. If the empirical gradient errors are omitted, the difference between the two squared norms is $-4\eta^2\,\mathrm{Re}(s\bar{d})/r^4$. Using (62) in the remaining terms of the direct norm expansion gives

$$\left|\left\|\widehat{\mathbf{h}}^+\right\|_2^2 - \left\|\widehat{\mathbf{x}}^+\right\|_2^2\right| \le C\eta^2|d| + C\eta\delta\sqrt{\mathcal{E}}.$$

The raw norms are bounded above and below by numerical constants, so the calculation leading to (57) yields

$$|c_+ - 1| \le C\eta^2|d| + C\eta\delta\sqrt{\mathcal{E}}. \tag{63}$$

Substituting this estimate into the two exact coordinate-change identities in the proof of Lemma 8 gives

$$s^+ = (1 - \eta + \eta/r^2)s + \zeta_s, \tag{64}$$

$$d^+ = (1 - \eta - \eta/r^2)d + \zeta_d, \tag{65}$$

$$\mathbf{h}_\perp^+ = (1-\eta)\mathbf{h}_\perp + \boldsymbol{\zeta}_\mathbf{h}, \qquad \mathbf{x}_\perp^+ = (1-\eta)\mathbf{x}_\perp + \boldsymbol{\zeta}_\mathbf{x}, \tag{66}$$

where

$$|\zeta_s| \le C\eta\delta\sqrt{\mathcal{E}} + C\eta^2 W, \tag{67}$$

$$|\zeta_d| + \|\boldsymbol{\zeta}_\mathbf{h}\|_2 + \|\boldsymbol{\zeta}_\mathbf{x}\|_2 \le C(\eta\delta + \eta^2)\sqrt{\mathcal{E}}. \tag{68}$$

For example, the term linear in $c_+ - 1$ in the common signal component is $(c_+^{-1} - c_+)\widehat{d}/2$, which accounts for the coefficient W in (67); in the other three components it is bounded by $C|c_+ - 1|$. This proves the two displayed remainder bounds without discarding their dependence on the current residual.

For the common signal component, let $\bar{s}^+$ denote the value obtained by omitting the empirical gradient error. It satisfies the exact identity

$$\bar{s}^+ - 1 = e - \eta\frac{s}{r^2}\{(s+1)e + W\}. \tag{69}$$

Under (60), the coefficient $s(s+1)/r^2$ lies between two positive numerical constants. Hence

$$|\bar{s}^+ - 1| \le (1 - c\eta)|e| + C\eta W. \tag{70}$$

Also, the three population multipliers in (65)–(66) have magnitude at most $1 - c\eta$. Squaring these estimates and using $2ab \le \tau a^2 + \tau^{-1}b^2$ with $\tau = c\eta$, we obtain

$$\begin{aligned}(|s^+| - 1)^2 &\le (1 - c\eta)e^2 + C\eta W^2 + C\eta(\delta^2 + \eta^2)\mathcal{E},\\ W^+ &\le (1 - c\eta)W + C\eta(\delta^2 + \eta^2)\mathcal{E}.\end{aligned}$$

Here $||s^+| - 1| \le |s^+ - 1|$ was used in the first line. Since $W^2 \le c_0^2 W$, choosing c_0, c_1 sufficiently small and adding the two inequalities proves (58).

Finally, choose $c = e^{-i\arg s}$ in (6). Then $p = s + d$, $q = s - d$, and orthogonality in (24) gives

$$\|\mathbf{h}/\bar{c} - \mathbf{h}^\star\|_2^2 + \|c\mathbf{x} - \mathbf{x}^\star\|_2^2 \le C\{(|s| - 1)^2 + |d|^2 + U + V\} = C\mathcal{E},$$

which is (59). □

The next lemma combines the preceding one-step estimates to show that the finite-sample trajectory reaches the local convergence region during Stage I.

Lemma 10. *Let $L \geq \log K \geq 2$ and suppose*

$$\frac{c_0}{K \log K} \leq S_0 \leq \frac{C_0 \log K}{K}, \quad D_0 \leq \frac{C_0 \log K}{K}, \quad c_0 \leq E_0 \leq W_0 \leq C_0, \qquad r_0 = 1. \tag{71}$$

Assume the empirical gradient approximation of Lemma 7, with $0 < \eta \leq cL^{-5}$ and $0 < \delta \leq c/(\sqrt{K}L^5)$. For $\gamma = c/(\sqrt{K}L^{3/2})$, there is $T_\gamma \leq C\eta^{-1}L$ such that

$$\operatorname{dist}(\mathbf{z}^{T_\gamma}, \mathbf{z}^\star) \leq C\gamma. \tag{72}$$

Moreover, $c/L^2 \leq r_t^2 \leq C$ for all $t \leq T_\gamma$.

Proof. This proof follows the three mechanisms identified in the population calculation. First, the undesired energy contracts while the common vector norm stays bounded away from zero. Second, the useful and undesired energies cross, after which the useful component grows. Third, the iterates enter the local convergence region, where Lemma 9 gives linear contraction. Define $\omega := C\delta^2 L^4$ and $\widetilde{W}_t := W_t + \omega$. The parameter conditions ensure $\omega \leq c\gamma^2/L^2$ and $\delta L\sqrt{K \log K} \leq c/L^3$.

Step 1: bounds after one update. We begin by separating the mismatch energy from the two orthogonal energies. Square (47)–(48), use (49)–(50), and apply $2ab \leq \varepsilon a^2 + \varepsilon^{-1}b^2$. We invoke Lemma 8 with $\ell = L^2$. As long as $L^{-2} \leq r_t^2 \leq 4$, the parameter assumptions give $\eta/r_t^2 \leq c/L^3$ and $\eta^2/r_t^4 \leq c\eta/L$. More explicitly, put

$$a_t = 1 + \eta(r_t^{-2} - 1), \qquad b_t = 1 - \eta(1 + r_t^{-2}).$$

The first two lines of Lemma 8 imply

$$\left|S_{t+1} - a_t^2 S_t\right| \leq 2a_t\sqrt{S_t}\,|\xi_{s,t}| + |\xi_{s,t}|^2.$$

Substitution of (49), followed by division by S_t, gives the first two inequalities below.

We next keep the mismatch energy D_t separate from the perpendicular energy $E_t = (U_t + V_t)/2$. Applying Young's inequality with parameter η to (47)–(48) gives

$$E_{t+1} \leq (1 - c\eta)E_t + \frac{C\eta}{L^2}D_t + C\eta\delta^2 L^2, \tag{73}$$

$$E_{t+1} \geq (1 - C\eta)E_t - \frac{C\eta}{L^2}D_t - C\eta\delta^2 L^2, \tag{74}$$

$$D_{t+1} \leq \{1 - c\eta(1 + r_t^{-2})\}D_t + \frac{C\eta}{L^2}E_t + C\eta\delta^2 L^2. \tag{75}$$

To see the powers of L, note that division by the Young parameter turns the squared remainder into

$$C\eta\delta^2/r_t^2 + C\eta^3 r_t^{-8}(D_t + E_t).$$

Here $\eta^2 r_t^{-8} = (\eta/r_t^4)^2 \leq c/L^2$ and $r_t^{-2} \leq L^2$. In (75), the unperturbed multiplier b_t^2 supplies the stronger negative term $-c\eta(1 + r_t^{-2})D_t$. These observations give all three displayed inequalities directly.

Set $\nu = C(\log K/K + L^{-2})$. The next induction shows that the mismatch energy remains a small fraction of the orthogonal energy. The added quantity ω absorbs the state-independent finite-sample terms. The initialization obeys $D_0 \leq \nu E_0/4$ after increasing C. Equations (73)–(75), and the fact that $\omega \geq C\delta^2 L^2$, show inductively that

$$D_t + \omega \leq \nu(E_t + \omega) \tag{76}$$

as long as $L^{-2} \leq r_t^2 \leq 4$. Indeed, if equality is approached, the three preceding recurrences, with their additive terms absorbed by ω, have the form

$$\overline{E}_{t+1} \geq (1 - 2\eta - C\eta/L)\overline{E}_t - C\eta L^{-2}\overline{D}_t,$$
$$\overline{D}_{t+1} \leq (1 - 3\eta + C\eta/L)\overline{D}_t + C\eta L^{-2}\overline{E}_t,$$

where $\overline{E}_t = E_t + \omega$ and $\overline{D}_t = D_t + \omega$; the coefficient 3 can be replaced by any fixed number strictly larger than 2 by reducing the step-size constant, since $1 + r_t^{-2} \geq 3/2$. Therefore, whenever $\overline{D}_t \leq \nu \overline{E}_t$,

$$\overline{D}_{t+1} - \nu \overline{E}_{t+1} \leq \left\{ -c\eta\nu + \frac{C\eta}{L^2} \right\} \overline{E}_t \leq 0,$$

where the last inequality follows from the definition of ν and a sufficiently large constant C. This proves (76) by induction. Since $\nu = o(1)$, (76) and (73)–(74) imply

$$\left(1 - 2\eta - \frac{C\eta}{L}\right)(E_t + \omega) \leq E_{t+1} + \omega \leq \left(1 - 2\eta + \frac{C\eta}{L}\right)(E_t + \omega). \tag{77}$$

Because $W_t = E_t + D_t$, (76) makes $\widetilde{W}_t$ comparable, up to numerical constants, with $E_t + \omega$. We therefore obtain

$$S_{t+1} \geq \left[1 + 2\eta(r_t^{-2} - 1) - C\eta^2 r_t^{-4} - C\eta \frac{\delta}{r_t \sqrt{S_t}}\right] S_t, \tag{78}$$

$$S_{t+1} \leq \left[1 + 2\eta(r_t^{-2} - 1) + C\eta^2 r_t^{-4} + C\eta \frac{\delta}{r_t \sqrt{S_t}}\right] S_t, \tag{79}$$

$$\widetilde{W}_{t+1} \leq \left(1 - 2\eta + \frac{C\eta}{L}\right) \widetilde{W}_t \leq (1 - c\eta)\widetilde{W}_t, \tag{80}$$

The last line follows directly by adding (73) and (75); the $C\eta L^{-2}(D_t + E_t)$ terms are absorbed by the negative leading term, and the remaining finite-sample term is absorbed by ω. Equation (51) also gives

$$r_{t+1}^2 \leq r_t^2 + C\eta(1 - r_t^2)_+ + C\eta\delta + C\eta^2 / r_t^2, \tag{81}$$

Since the argument below stops within $C\eta^{-1}L$ steps and $\delta L + \eta L^3 = o(1)$, an induction using this inequality gives $r_t^2 \leq 2$ throughout the time interval under consideration.

We claim that $S_t \geq S_0/2$ until the useful energy crosses the undesired energy. Indeed, under this claim and $r_t^2 \geq L^{-2}$,

$$\frac{\delta}{r_t \sqrt{S_t}} \leq C\delta L \sqrt{K \log K} \leq \frac{c}{L^3}.$$

Whenever $r_t^2 \leq 1 - c/L$, the positive leading term in (78) dominates both error terms. While $r_t^2 > 1 - c/L$, the same inequality gives $S_{t+1} \geq (1 - C\eta/L)S_t$, while (79) gives $S_{t+1} \leq (1 + C\eta/L)S_t$. Hence during the first $C\eta^{-1}$ steps of this interval, $S_t \leq e^{C/L} S_0 = o(1)$. At the same time, (80) gives $\widetilde{W}_t \leq e^{-c\eta t}\widetilde{W}_0$. Thus the inequality $r_t^2 = S_t + W_t > 1 - c/L$ holds for at most $C\eta^{-1}$ steps. Multiplying the one-step lower bounds over this interval gives $S_t \geq e^{-C/L} S_0 \geq S_0/2$.

Step 2: the useful energy catches up with the undesired energy. We now show that the useful energy reaches the scale of the undesired energy before r_t^2 can fall below L^{-2}. Let $z_t = S_t / \widetilde{W}_t$ and $\Phi(z) = \log z + z$. The inequalities $\log(1 + u) \geq u - Cu^2$ and $\log(1 - u) \leq -u$ (valid here because $\eta / r_t^2 \leq c/L^3$) applied to (78) and the first inequality in (80) give

$$\log \frac{z_{t+1}}{z_t} \geq \frac{2\eta}{r_t^2} - \frac{C\eta}{L}. \tag{82}$$

The right-hand side is positive after reducing the numerical constants. Since $e^u - 1 \geq u$ for $u \geq 0$,

$$\Phi(z_{t+1}) - \Phi(z_t) = \log(z_{t+1}/z_t) + (z_{t+1} - z_t) \geq (1 + z_t) \log(z_{t+1}/z_t).$$

Finally, when $z_t < 1$ and $W_t > L^{-2}$, the added term satisfies $\omega \ll W_t$, and

$$\frac{1 + z_t}{r_t^2} = \frac{\widetilde{W}_t + S_t}{\widetilde{W}_t (S_t + W_t)} \geq \frac{c}{\widetilde{W}_t}.$$

Combining the last three displays gives

$$\Phi(z_{t+1}) - \Phi(z_t) \geq c \frac{\eta}{\widetilde{W}_t} - C \frac{\eta}{L}. \tag{83}$$

as long as $z_t < 1$ and $W_t > L^{-2}$.

We now use a stopping argument, which also verifies that $r_t^2 \geq L^{-2}$ while the preceding estimates are applied. Let $\tau = \inf\{t : z_t \geq 1 \text{ or } W_t \leq L^{-2}\}$. For $t < \tau$, $r_t^2 = S_t + W_t > L^{-2}$, so all preceding estimates are valid. Suppose for contradiction that the second event occurs first. Equations (76) and (77) imply that every decrease of $\widetilde{W}_t$ by a factor of two requires at least c/η iterations. Indeed, up to a factor $1 + \mathcal{O}(\nu)$, $\widetilde{W}_t = E_t + \omega$, and the latter can decrease by at most the factor $1 - C\eta$ in one step. Applying this observation to the last interval in which $\widetilde{W}_t$ decreases by a factor of two before reaching L^{-2} gives

$$\sum_{t<\tau} \frac{\eta}{\widetilde{W}_t} \geq cL^2. \tag{84}$$

Throughout that range each of the c/η summands is at least $c\eta L^2$. On the other hand, (71) implies $\Phi(z_0) \geq -CL$. Equation (80) also gives $\tau \leq C\eta^{-1} \log L$ under the supposition that the norm threshold occurs first. Thus the accumulated negative term in (83) contributes $\mathcal{O}(\log L/L)$. Summing (83) and using (84) therefore forces $\Phi(z_t) \geq \Phi(1)$ strictly before W_t reaches L^{-2}, a contradiction. Thus the useful energy crosses first. Moreover, $\widetilde{W}_t \leq Ce^{-c\eta t} + \omega$ before crossing, so it does so by

$$T_\times \leq C\eta^{-1} \log L. \tag{85}$$

At crossing, $S_{T_\times} \geq c/L^2$. If subsequently S_t approaches a sufficiently small constant multiple of L^{-2}, then $r_t^2 \leq 1 - c/L^2$ and the positive drift in (78) dominates its two error terms. Hence $S_t \geq c/L^2$ thereafter. Consequently $r_t^2 = S_t + W_t \geq c/L^2$ throughout the induction, validating the assumed range of r_t in Step 1.

Step 3: entry into the local convergence region. After the crossing, the useful energy grows while the undesired energy keeps contracting. We first show that the useful energy reaches a fixed positive constant and then show that the iterate enters the local convergence region. At the first crossing, $S_{T_\times} \geq \widetilde{W}_{T_\times}$. The lower bound on the common signal component and the strict contraction of $\widetilde{W}_t$ imply that $W_t \leq 2S_t$ until S_t reaches $1/4$. Indeed, suppose this inequality fails for the first time. At the preceding iterate one has $W_t \leq 2S_t$, while the calculation below gives $S_{t+1} \geq S_t + c\eta$ and (80) gives $W_{t+1} \leq W_t$; hence $W_{t+1} < 2S_{t+1}$, a contradiction. On this interval, $r_t^2 = S_t + W_t \leq 3S_t \leq 3/4$, and (78) gives

$$S_{t+1} - S_t \geq 2\eta S_t(r_t^{-2} - 1) - C\eta^2 S_t/r_t^4 - C\eta\delta\sqrt{S_t}/r_t \geq c\eta.$$

The final inequality uses $S_t/r_t^2 \geq 1/3$, $\eta/r_t^2 \leq c/L^3$, and $\delta/(r_t\sqrt{S_t}) \leq c/L^3$. Thus S_t reaches $1/4$ within $C\eta^{-1}$ further steps.

We next derive a recurrence for the distance between the useful energy and one. If $\overline{S}_{t+1}$ denotes the value obtained by omitting the empirical gradient error, then

$$\overline{S}_{t+1} - S_t = 2\eta \frac{S_t}{S_t + W_t}(1 - S_t - W_t) + \eta^2(r_t^{-2} - 1)^2 S_t. \tag{86}$$

For $1/4 \leq S_t \leq 3/2$ and $W_t \leq 2S_t$, the first term in (86) contracts $1 - S_t$ toward zero, with W_t as an additive perturbation. The second term is $\mathcal{O}(\eta^2)$ in this region. Using (49) to compare S_{t+1} and $\overline{S}_{t+1}$ therefore gives

$$|1 - S_{t+1}| \leq (1 - c\eta)|1 - S_t| + C\eta W_t + C\eta\delta + C\eta^2. \tag{87}$$

At $S = 1/4$ and $S = 3/2$, the same formula moves S toward the interior of the interval. Thus $1/4 \leq S_t \leq 3/2$ is preserved after reducing the step-size constant. Meanwhile, iteration of (80) yields

$$\widetilde{W}_t \leq e^{-c\eta(t-T_\times)}\widetilde{W}_{T_\times}. \tag{88}$$

Iterating (87) gives the explicit bound

$$\begin{aligned} |1 - S_t| &\leq e^{-c\eta(t-T_\times)}|1 - S_{T_\times}| + C\sum_{k=T_\times}^{t-1} e^{-c\eta(t-1-k)}\{\eta W_k + \eta\delta + \eta^2\} \\ &\leq Ce^{-c\eta(t-T_\times)} + C(\delta + \eta). \end{aligned} \tag{89}$$

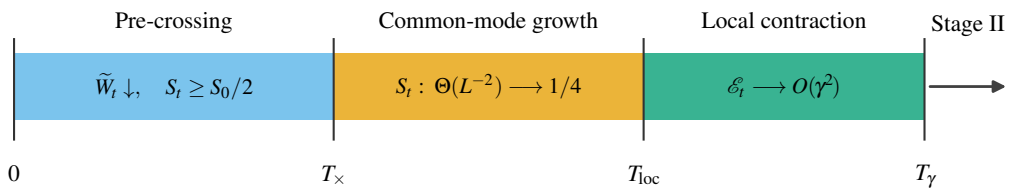


Figure 7: The three intervals in the Stage-I proof.

Choose the numerical constants in the assumptions of Lemma 10 small enough relative to c_0 in Lemma 9. Equations (88)–(89) then show that at some $T_{\mathrm{loc}} \leq T_\times + C\eta^{-1}$ one has

$$(|s_{T_{\mathrm{loc}}}| - 1)^2 + W_{T_{\mathrm{loc}}} \leq c_0^2. \tag{90}$$

Starting at T_{loc}, apply Lemma 9 repeatedly. For $N = \lceil (C/\eta)\log(c_0/\gamma) \rceil \leq C\eta^{-1}L$ and $T_\gamma = T_{\mathrm{loc}} + N$, it gives

$$W_{T_\gamma} \leq \mathcal{E}_{T_\gamma} \leq C\gamma^2, \qquad (|s_{T_\gamma}| - 1)^2 \leq C\gamma^2. \tag{91}$$

Let $\theta = \arg s_{T_\gamma}$ and take $c = e^{-i\theta}$ in (6). Since $p = s + d$ and $q = s - d$,

$$\|\mathbf{h}/\overline{c} - \mathbf{h}^\star\|_2^2 + \|c\mathbf{x} - \mathbf{x}^\star\|_2^2 \leq C\{(\sqrt{S} - 1)^2 + W\}.$$

Equation (72) follows from (59). Before the crossing, the stopping argument gives $r_t^2 \geq L^{-2}$. After the crossing, the lower-bound argument for S_t gives $r_t^2 \geq c/L^2$, and after local entry (60) gives $r_t^2 \geq 1/2$. Together with the upper bound proved in Step 1, this proves the asserted bounds on r_t. □

Remark 11 (The substages of Stage I)**.** *The proof divides the trajectory into three intervals. Before $T_\times$, the undesired energy contracts while the common signal is protected from vanishing. Between $T_\times$ and T_{loc}, the common signal grows to constant scale. Between T_{loc} and T_γ, the local energy contracts to the target accuracy. These three times mark the transitions between the estimates used in the proof. Sample dependence is handled uniformly over all three intervals by the RIP event in Lemma 7.*

E Random initialization and the Stage I conclusion

We now verify that the random initialization satisfies the deterministic conditions used in the Stage-I lemma. The two initial truth correlations are typically of order $K^{-1/2}$, while the orthogonal energy is of constant order. We bound the probability that their average is unusually close to zero and use a Gaussian tail bound for the remaining initial quantities.

Lemma 12. *Under the independent random initialization of Section 2, with probability at least $1 - C/\log K - K^{-10}$,*

$$\frac{c}{K\log K} \leq S_0 \leq \frac{C\log K}{K}, \qquad D_0 \leq \frac{C\log K}{K}, \qquad c \leq E_0 \leq W_0 \leq C. \tag{92}$$

Proof. By the initialization in Section 2, each initial vector is uniformly distributed on the complex unit sphere $\mathbb{S}_{\mathbb{C}}^{K-1}$. Hence, if $\mathbf{u}$ denotes either initial vector and $\mathbf{v}$ is fixed and unit-norm, then $Z = \mathbf{v}^{\mathsf{H}}\mathbf{u}$ has density

$$f_Z(z) = \frac{K-1}{\pi}(1 - |z|^2)^{K-2}\mathbf{1}_{\{|z|\leq 1\}}.$$

Thus $\|f_Z\|_\infty \leq K/\pi$ and $\mathbb{P}(|Z| \geq t) = (1-t^2)^{K-1} \leq e^{-(K-1)t^2}$. The variables p_0 and q_0 are independent copies of Z. The density of $s_0 = (p_0 + q_0)/2$ is bounded by $4K/\pi$, whence

$$\mathbb{P}\left(|s_0| \leq \frac{c}{\sqrt{K\log K}}\right) \leq \frac{C}{\log K}.$$

The same tail bound gives $|p_0| + |q_0| \le C\sqrt{\log K/K}$ except with probability K^{-10}. Finally,

$$E_0 = 1 - \frac{|p_0|^2 + |q_0|^2}{2}, \qquad D_0 = \frac{|p_0 - q_0|^2}{4}, \qquad W_0 = E_0 + D_0,$$

which proves (92). □

Proof of the Stage I claims in Theorem 1. We combine the uniform RIP event, the Gaussian norm event, and the initialization event. Each event is uniform over the trajectory, so no union bound over the individual iterations is needed. Set $\delta = c_1/(\sqrt{K}\log^5 m)$. The sample-size condition in Theorem 1, with a sufficiently large constant, implies the lower bound required by Lemma 6. Hence Lemmas 6 and 7 hold with failure probability at most K^{-10}. Intersect this event with the Gaussian norm event (46) and the event in Lemma 12. Their total failure probability is at most $C/\log K + CK^{-10} + Cme^{-cK}$.

Take $L = \log m$. The initialization and parameter conditions of Lemma 10 follow from (92), the step-size condition in Theorem 1, and our choice of δ. Lemma 10 proves that $T_\gamma \le C\eta^{-1}\log m$, that the quotient distance at T_γ is at most $C\gamma$, and that $c\log^{-2} m \le r_t^2 \le C$ throughout Stage I.

Align the pair at T_γ as in (6). On the Gaussian norm event (46), the distance bound just proved gives

$$\max_j |\mathbf{a}_j^{\mathsf{H}}(\mathbf{x}^{T_\gamma} - \mathbf{x}^\star)| \le 3\sqrt{K}\, C_0\gamma \le \frac{C_0}{\log^{3/2} m}.$$

Similarly, the Fourier-frame identity and the incoherence assumption on the ground truth from Section 3, together with the distance bound, give

$$\max_j |\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^{T_\gamma}| \le \frac{\mu}{\sqrt{m}} + \sqrt{\frac{K}{m}}\, C_0\gamma \le \frac{\mu + C_0\log^{-3/2} m}{\sqrt{m}}.$$

These are the two incoherence bounds needed to start Stage II. The distance at the end of Stage I is also smaller than the $\mathcal{O}(\log^{-2} m)$ radius in the local restricted-strong-convexity result of Ma et al. (2018). □

F Stage II: local linear convergence

Stage I ends inside the local region where the factor vectors are already close to the truth. We complete the proof by applying the local contraction result for blind deconvolution in (Ma et al., 2018, Lemma 15). That result holds uniformly over all nearby points satisfying its incoherence conditions and therefore applies to the point reached at the end of Stage I. We verify the hypotheses at the first Stage-II iterate and then verify them inductively at every later iterate. To keep this alignment distinct from the equal-norm coordinates used in Stage I, let $\alpha = \alpha(\mathbf{h}, \mathbf{x})$ attain the minimum in the quotient distance and define $\mathbf{h}_{\mathrm{a}} := \mathbf{h}/\overline{\alpha}$ and $\mathbf{x}_{\mathrm{a}} := \alpha\mathbf{x}$. These are the variables denoted by $(\widetilde{\mathbf{h}}, \widetilde{\mathbf{x}})$ in Ma et al. (2018).

Proof of the Stage II claim in Theorem 1. We first check the local contraction lemma at the point where Stage I ends. Because the distance decreases after every update, the same bounds remain valid at all later iterations. Let $\mathcal{E}_{\mathrm{BD,loc}}$ denote the high-probability event in (Ma et al., 2018, Lemma 15). The lemma states that, provided $m \ge C\mu^2 K\log^9 m$ and η is a sufficiently small numerical constant, every iterate satisfying

$$\operatorname{dist}(\mathbf{z}, \mathbf{z}^\star) \le \xi, \tag{93}$$

$$\max_j |\mathbf{a}_j^{\mathsf{H}}(\mathbf{x}_{\mathrm{a}} - \mathbf{x}^\star)| \le C_3\log^{-3/2} m, \tag{94}$$

$$\max_j |\mathbf{b}_j^{\mathsf{H}}\mathbf{h}_{\mathrm{a}}| \le C_4\mu m^{-1/2}\log^2 m \tag{95}$$

obeys the one-step contraction

$$\operatorname{dist}(\mathbf{z}^+, \mathbf{z}^\star) \le (1 - \eta/16)\operatorname{dist}(\mathbf{z}, \mathbf{z}^\star). \tag{96}$$

Here $\xi \ll \log^{-2} m$. The event $\mathcal{E}_{\mathrm{BD,loc}}$ is independent of the iteration number and has probability at least $1 - Cm^{-10} - Ce^{-cK} \log m$.

We now verify these assumptions. First, the Fourier-frame assumption gives $\max_j |\mathbf{b}_j^{\mathsf{H}} \mathbf{h}^\star| \le \sqrt{K/m}$, so the minimal incoherence parameter satisfies $1 \le \mu \le \sqrt{K}$. Therefore, using $K, \log m \ge 1$, we have $\mu^2 K \log^9 m \le K^2 \log^9 m \le CK^3 \log^{11} m \le m$. Also, our step size is smaller than a sufficiently small numerical constant. Intersect $\mathcal{E}_{\mathrm{BD,loc}}$ with the Stage I event and the Gaussian norm event (46).

Set $d_t = \mathrm{dist}(\mathbf{z}^t, \mathbf{z}^\star)$. At $t = T_\gamma$, the Stage I conclusion gives $d_t \le C_0 \gamma$. Since $\log m \le cK$, after reducing the constant in γ, $C_0 \gamma = C_0 c/(\sqrt{K} \log^{3/2} m) \ll 1/\log^2 m$, which verifies (93). By the definition of the quotient distance, each aligned vector error of $(\mathbf{h}_{\mathrm{a}}^t, \mathbf{x}_{\mathrm{a}}^t)$ is at most d_t. Thus (46) gives

$$\max_j |\mathbf{a}_j^{\mathsf{H}} (\mathbf{x}_{\mathrm{a}}^t - \mathbf{x}^\star)| \le 3\sqrt{K}\, d_t \le 3C_0 \sqrt{K}\, \gamma \le C_3 \log^{-3/2} m, \tag{97}$$

and the Fourier-frame identity and the incoherence assumption on the ground truth give

$$\max_j |\mathbf{b}_j^{\mathsf{H}} \mathbf{h}_{\mathrm{a}}^t| \le \frac{\mu}{\sqrt{m}} + \sqrt{\frac{K}{m}}\, d_t \le \frac{\mu + C_0 \log^{-3/2} m}{\sqrt{m}} \le C_4 \frac{\mu}{\sqrt{m}} \log^2 m. \tag{98}$$

Here we used $\mu \ge 1$ in the last step. Hence all three hypotheses of (Ma et al., 2018, Lemma 15) hold at T_γ, and (96) applies.

The verification is inductive. If it holds at time t, then (96) gives $d_{t+1} \le d_t$. Replacing d_t by the smaller d_{t+1} in (97)–(98) verifies (93)–(95) at time $t+1$. Therefore (96) can be applied for every $t \ge T_\gamma$. Iteration proves the contraction bound (7).

Finally, selecting equal-norm representatives in the Stage-I proof causes no change in this argument. The normalized update is equivariant under $(\mathbf{h}, \mathbf{x}) \mapsto (\mathbf{h}/\bar{c}, c\mathbf{x})$, and both the distance and the conditions above are formulated after optimal alignment. Thus the equal-norm representatives used here and the aligned representatives in the cited lemma produce the same sequence of recovered matrices. Combining this Stage-II argument with the preceding Stage-I proof completes Theorem 1. The additional failure probability of $\mathcal{E}_{\mathrm{BD,loc}}$ is absorbed by the last two failure terms in the probability stated there. □

G LOCAL CONTRACTION FOR THE BDD EXTENSION

This section gives the local-convergence argument associated with Theorem 3. The proof follows the local blind-deconvolution analysis in Ma et al. (2018). We separate the new BDD calculation from the steps inherited from that analysis. The new part is the population restricted strong convexity proved in the next subsection.

We first describe the local region used in Theorem 3. For each source, choose α_i that attains the minimum in the quotient distance and write $\mathbf{h}_{i,\mathrm{a}} = \mathbf{h}_i/\overline{\alpha_i}$ and $\mathbf{x}_{i,\mathrm{a}} = \alpha_i \mathbf{x}_i$. Stack each pair and its complex conjugate in the order $(\mathbf{h}_{i,\mathrm{a}}, \mathbf{x}_{i,\mathrm{a}}, \overline{\mathbf{h}_{i,\mathrm{a}}}, \overline{\mathbf{x}_{i,\mathrm{a}}})$ and define

$$\begin{aligned} \mathbf{P} &:= \mathrm{Diag}(\mathbf{P}_1, \dots, \mathbf{P}_n), \\ \mathbf{P}_i &:= \mathrm{Diag}(\gamma_{i1}\mathbf{I}_K, \gamma_{i2}\mathbf{I}_K, \gamma_{i1}\mathbf{I}_K, \gamma_{i2}\mathbf{I}_K)\,, \\ \gamma_{i1} &:= \|\mathbf{x}_{i,\mathrm{a}}\|_2^{-2}\,, \qquad \gamma_{i2} := \|\mathbf{h}_{i,\mathrm{a}}\|_2^{-2}\,. \end{aligned} \tag{99}$$

This block-diagonal matrix is the preconditioner induced by the two denominators in the normalized update. Let $\mathbf{H}$ denote the augmented Wirtinger Hessian of the empirical BDD loss at the same aligned state. Let $\mathbf{u}$ stack the differences between these aligned factor vectors and the ground-truth factor vectors, together with their conjugates. Optimal alignment removes the ambiguous scaling direction of each source. We impose the lower Hessian bound on these aligned error vectors, as stated precisely in Lemma 13.

Let $\mathcal{E}_{\mathrm{BDD,loc}}$ be the high-probability event on which the finite-sample BDD bounds hold. Let $\mathcal{R}$ be a neighborhood of the ground-truth equivalence class in the aligned factor-pair space, chosen so that every aligned state in $\mathcal{R}$ satisfies

$$\|\mathbf{PH}\|_2 \le 3, \qquad \mathbf{u}^{\mathsf{H}} (\mathbf{PH} + \mathbf{HP})\, \mathbf{u} \ge \frac{1}{4} \|\mathbf{u}\|_2^2 \tag{100}$$

for every aligned error vector $\mathbf{u}$. The radius of $\mathcal{R}$ is small enough to control the error caused by re-aligning after an update. On $\mathcal{E}_{\mathrm{BDD,loc}}$, a leave-one-out induction establishes the required incoherence bounds and keeps the empirical iterates in this region. Thus the first inequality controls the largest local curvature, the second controls the curvature toward the ground truth, and the induction ensures that both remain valid along the trajectory.

Proof of Theorem 3. We divide the proof into three steps. We first establish the local Hessian bounds, then prove one-step contraction, and finally explain why the local conditions remain valid along the trajectory.

Step 1: Hessian bounds. Fix an aligned state in $\mathcal{R}$ and form $\mathbf{P}$ and $\mathbf{H}$ at that state. On the event $\mathcal{E}_{\mathrm{BDD,loc}}$, the assumptions (100) give, for every aligned error vector $\mathbf{v}$,

$$\|\mathbf{PHv}\|_2 \le 3\,\|\mathbf{v}\|_2\,, \qquad 2\,\mathrm{Re}\langle \mathbf{v}, \mathbf{PHv}\rangle \ge \frac{1}{4}\,\|\mathbf{v}\|_2^2\,. \tag{101}$$

The second inequality is the restricted strong convexity bound. For the population loss at the ground truth, Lemma 13 proves this bound with a numerical constant independent of κ. The Hessian block for source i contributes σ_i^2, while the normalization in its gradient update contributes σ_i^{-2}. The empirical version in (101) follows by uniformly bounding the difference between the empirical and population Hessians over nearby points that satisfy the required incoherence conditions.

Step 2: one-step contraction. Let $\mathbf{Z}^+$ be the simultaneous normalized update of $\mathbf{Z}$, and let $\mathbf{u}$ and $\mathbf{u}^+$ be the aligned errors before and after this update. Apply the mean-value form of Taylor's theorem on the line segment between the aligned iterate and the ground truth to obtain

$$\mathbf{u}^+ = (\mathbf{I} - \eta \mathbf{P}_\tau \mathbf{H}_\tau)\,\mathbf{u} + \mathbf{e}_{\mathrm{loc}}, \qquad 0 \le \tau \le 1, \tag{102}$$

where $\mathbf{P}_\tau$ and $\mathbf{H}_\tau$ are evaluated at the same point on this segment. The remainder $\mathbf{e}_{\mathrm{loc}}$ contains the variation of the preconditioner along the segment and the change caused by re-aligning the updated factor vectors. Since the gradient vanishes at every pair of factor vectors representing the ground truth, the first contribution is quadratic in the local error. The alignment estimate for the quotient distance in Ma et al. (2018) bounds this term by

$$\|\mathbf{e}_{\mathrm{loc}}\|_2 \le c_{\mathrm{al}}\eta\,\|\mathbf{u}\|_2^2\,. \tag{103}$$

The local radius is chosen small enough that this quadratic remainder is at most $(\eta/64)\,\|\mathbf{u}\|_2$.

Ignoring this quadratic remainder for the moment and expanding the squared norm gives

$$\begin{aligned}\|(\mathbf{I} - \eta\mathbf{P}_\tau\mathbf{H}_\tau)\mathbf{u}\|_2^2 &= \|\mathbf{u}\|_2^2 - 2\eta\,\mathrm{Re}\langle \mathbf{u}, \mathbf{P}_\tau\mathbf{H}_\tau\mathbf{u}\rangle + \eta^2\,\|\mathbf{P}_\tau\mathbf{H}_\tau\mathbf{u}\|_2^2 \\ &\le \Big(1 - \frac{\eta}{4} + 9\eta^2\Big)\,\|\mathbf{u}\|_2^2\,.\end{aligned} \tag{104}$$

Taking $\eta \le 1/144$ makes the parenthesis at most $1 - 3\eta/16$. Since $\sqrt{1-a} \le 1 - a/2$ for $0 \le a \le 1$, the linear part contracts by at least $3\eta/32$. Combining this bound with (103) yields

$$\mathrm{dist}_{\mathrm{BDD}}(\mathbf{Z}^+, \mathbf{Z}^\star) \le \Big(1 - \frac{\eta}{16}\Big)\,\mathrm{dist}_{\mathrm{BDD}}(\mathbf{Z}, \mathbf{Z}^\star). \tag{105}$$

The key point is that the upper and lower Hessian bounds are numerical constants. Lemma 13 shows that the source strengths cancel after normalization, yielding a contraction coefficient independent of κ.

Step 3: staying in the local region. It remains to verify that the hypotheses above hold for every iterate. For each sample index ℓ, construct a leave-one-out BDD trajectory by deleting all sensing vectors with index ℓ, namely $(\mathbf{a}_{1\ell}, \ldots, \mathbf{a}_{n\ell})$, from the empirical loss. The resulting iterate is independent of these deleted vectors. Conditional concentration therefore controls the residual for sample ℓ and the Fourier incoherence terms evaluated at the leave-one-out iterate. A stability recursion comparing the leave-one-out and full trajectories, followed by a union bound over ℓ and the iteration index, transfers the same bounds to the actual trajectory. This is the leave-one-out induction used in the local BD proof of Ma et al. (2018); for BDD it also tracks the cross-source residuals.

Once these bounds are available, the difference between the empirical and population Hessians remains within the tolerance used in (101), and the local remainder remains within the tolerance in (103). Thus the assumptions needed at the next iterate follow from those at the current iterate, and (105) applies at every iteration. Iterating this inequality proves

$$\operatorname{dist}_{\mathrm{BDD}}(\mathbf{Z}^t, \mathbf{Z}^\star) \le \left(1 - \frac{\eta}{16}\right)^t \operatorname{dist}_{\mathrm{BDD}}(\mathbf{Z}^0, \mathbf{Z}^\star), \qquad t \ge 0. \tag{106}$$

This completes the local proof. The omitted leave-one-out and concentration estimates apply the corresponding arguments in Ma et al. (2018) to each source while also controlling the cross-source residuals. The next subsection proves the new population curvature bound that establishes the κ-independent local rate. □

G.1 Population restricted strong convexity for the BDD extension

This subsection proves the condition-number cancellation used in Section 5 and observed in Figure 5. We use symbols reserved for the BDD argument so that they remain distinct from the BD trajectory variables introduced earlier in the appendix.

Let n be the number of sources and let $\|\mathbf{h}_i^\star\|_2 = \|\mathbf{x}_i^\star\|_2 =: \sigma_i$ for $1 \le i \le n$. A common positive rescaling of the true and estimated factor vectors rescales the observations and preserves the relative normalized-gradient dynamics. We therefore normalize $\max_{1\le i\le n} \sigma_i = 1$ and define $\kappa := \max_i \sigma_i / \min_i \sigma_i$, so that $\kappa^{-1} \le \sigma_i \le 1$. This reciprocal common rescaling converts the experimental convention $\sigma_1 = 1, \sigma_2 = \kappa$ into the convention used here while preserving the relative errors and contraction behavior in Section 6.2.

For each source, consider two pairs of factor vectors $(\mathbf{h}_{i1}, \mathbf{x}_{i1})$ and $(\mathbf{h}_{i2}, \mathbf{x}_{i2})$. Choose their representatives so that $\iota = 1$ minimizes

$$\|\mathbf{h}_{i1}/\bar{\iota} - \mathbf{h}_{i2}\|_2^2 + \|\iota \mathbf{x}_{i1} - \mathbf{x}_{i2}\|_2^2, \tag{107}$$

and define

$$\Delta\mathbf{h}_i := \mathbf{h}_{i1} - \mathbf{h}_{i2}, \qquad \Delta\mathbf{x}_i := \mathbf{x}_{i1} - \mathbf{x}_{i2}, \qquad \mathbf{u}_i := \begin{bmatrix} \Delta\mathbf{h}_i \\ \Delta\mathbf{x}_i \\ \overline{\Delta\mathbf{h}_i} \\ \overline{\Delta\mathbf{x}_i} \end{bmatrix}, \tag{108}$$

with $\mathbf{u} := (\mathbf{u}_1^\mathsf{T}, \dots, \mathbf{u}_n^\mathsf{T})^\mathsf{T}$.

The next lemma proves the population curvature bound responsible for the condition-number-independent local contraction coefficient.

Lemma 13. *Assume that the aligned pairs of factor vectors satisfy*

$$\max_{\substack{1\le i\le n \\ \ell\in\{1,2\}}} \max\left\{\|\mathbf{h}_{i\ell} - \mathbf{h}_i^\star\|_2, \|\mathbf{x}_{i\ell} - \mathbf{x}_i^\star\|_2\right\} \le \vartheta. \tag{109}$$

Let $\mathbf{P} := \operatorname{Diag}(\mathbf{P}_1, \dots, \mathbf{P}_n)$ *be a diagonal preconditioner with blocks*

$$\mathbf{P}_i := \operatorname{Diag}\left(\gamma_{i1}\mathbf{I}_K, \gamma_{i2}\mathbf{I}_K, \gamma_{i1}\mathbf{I}_K, \gamma_{i2}\mathbf{I}_K\right), \qquad \max\left\{|\gamma_{i1} - \sigma_i^{-2}|, |\gamma_{i2} - \sigma_i^{-2}|\right\} \le \vartheta. \tag{110}$$

When the first pair is the current aligned iterate, the normalized update uses precisely $\gamma_{i1} = \|\mathbf{x}_{i,\mathrm{a}}\|_2^{-2}$ *and* $\gamma_{i2} = \|\mathbf{h}_{i,\mathrm{a}}\|_2^{-2}$*, as defined in (99). If* $0 < \vartheta \le \min\{1/2, 1/(39\kappa^2 + 2)\}$*, then the population Wirtinger Hessian* $\mathbf{H} := \mathbb{E}[\nabla^2 f_{\mathrm{BDD}}(\mathbf{Z}^\star)]$ *satisfies*

$$\|\mathbf{H}\|_2 = 2, \qquad \mathbf{u}^\mathsf{H}(\mathbf{P}\mathbf{H} + \mathbf{H}\mathbf{P})\,\mathbf{u} \ge \|\mathbf{u}\|_2^2. \tag{111}$$

Thus the RSC constant itself is numerical and independent of κ*, whereas this conservative sufficient neighborhood radius may shrink with* κ*.*

Proof. We first compute the population Hessian and its spectrum, and then lower-bound the symmetrized preconditioned quadratic form on the aligned directions in (108).

The sensing vectors are centered and independent across sources. Hence all cross-source Hessian blocks vanish after expectation. Using $\sum_{j=1}^m \mathbf{b}_j\mathbf{b}_j^\mathsf{H} = \mathbf{I}_K$, direct Wirtinger differentiation of (15) gives

$$\mathbf{H} = \operatorname{Diag}(\mathbf{H}_1, \dots, \mathbf{H}_n), \tag{112}$$

where

$$\mathbf{H}_i = \begin{bmatrix} \sigma_i^2\mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{h}_i^\star\mathbf{x}_i^{\star\mathsf{T}} \\ \mathbf{0} & \sigma_i^2\mathbf{I}_K & \mathbf{x}_i^\star\mathbf{h}_i^{\star\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & (\mathbf{x}_i^\star\mathbf{h}_i^{\star\mathsf{T}})^{\mathsf{H}} & \sigma_i^2\mathbf{I}_K & \mathbf{0} \\ (\mathbf{h}_i^\star\mathbf{x}_i^{\star\mathsf{T}})^{\mathsf{H}} & \mathbf{0} & \mathbf{0} & \sigma_i^2\mathbf{I}_K \end{bmatrix}. \tag{113}$$

The diagonal blocks follow from

$$\mathbb{E}|\mathbf{a}_{ij}^{\mathsf{H}}\mathbf{x}_i^\star|^2 = \sigma_i^2, \qquad \sum_{j=1}^{m} |\mathbf{b}_j^{\mathsf{H}}\mathbf{h}_i^\star|^2 = \sigma_i^2. \tag{114}$$

To diagonalize (113), introduce the four orthonormal vectors

$$\boldsymbol{\psi}_{i1} = \frac{1}{\sqrt{2}\sigma_i}\begin{bmatrix}\mathbf{h}_i^\star \\ \mathbf{0} \\ \mathbf{0} \\ \overline{\mathbf{x}_i^\star}\end{bmatrix}, \quad \boldsymbol{\psi}_{i2} = \frac{1}{\sqrt{2}\sigma_i}\begin{bmatrix}\mathbf{0} \\ \mathbf{x}_i^\star \\ \overline{\mathbf{h}_i^\star} \\ \mathbf{0}\end{bmatrix}, \quad \boldsymbol{\psi}_{i3} = \frac{1}{\sqrt{2}\sigma_i}\begin{bmatrix}\mathbf{h}_i^\star \\ \mathbf{0} \\ \mathbf{0} \\ -\overline{\mathbf{x}_i^\star}\end{bmatrix}, \quad \boldsymbol{\psi}_{i4} = \frac{1}{\sqrt{2}\sigma_i}\begin{bmatrix}\mathbf{0} \\ \mathbf{x}_i^\star \\ -\overline{\mathbf{h}_i^\star} \\ \mathbf{0}\end{bmatrix}.$$

A direct multiplication yields

$$\mathbf{H}_i = \sigma_i^2\left(\mathbf{I}_{4K} + \boldsymbol{\psi}_{i1}\boldsymbol{\psi}_{i1}^{\mathsf{H}} + \boldsymbol{\psi}_{i2}\boldsymbol{\psi}_{i2}^{\mathsf{H}} - \boldsymbol{\psi}_{i3}\boldsymbol{\psi}_{i3}^{\mathsf{H}} - \boldsymbol{\psi}_{i4}\boldsymbol{\psi}_{i4}^{\mathsf{H}}\right). \tag{115}$$

Therefore the eigenvalues of the ith block belong to $\{0, \sigma_i^2, 2\sigma_i^2\}$, and the normalization above gives $\|\mathbf{H}\|_2 = 2$.

Multiplying (113) by the ith preconditioner block gives

$$\mathbf{u}_i^{\mathsf{H}}\mathbf{P}_i\mathbf{H}_i\mathbf{u}_i = 2\gamma_{i1}\sigma_i^2\|\Delta\mathbf{h}_i\|_2^2 + 2\gamma_{i2}\sigma_i^2\|\Delta\mathbf{x}_i\|_2^2 + (\gamma_{i1}+\gamma_{i2})(\beta_i + \beta_i^*), \tag{116}$$

where the potentially negative contribution is determined by

$$\beta_i := [\Delta\mathbf{h}_i^{\mathsf{H}}\mathbf{h}_i^\star][\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_i^\star]. \tag{117}$$

Because $\|\mathbf{u}_i\|_2^2 = 2\|\Delta\mathbf{h}_i\|_2^2 + 2\|\Delta\mathbf{x}_i\|_2^2$, condition (110) implies

$$2\gamma_{i1}\sigma_i^2\|\Delta\mathbf{h}_i\|_2^2 + 2\gamma_{i2}\sigma_i^2\|\Delta\mathbf{x}_i\|_2^2 \geq (1-\vartheta)\|\mathbf{u}_i\|_2^2. \tag{118}$$

Furthermore, $\sigma_i^{-2} \leq \kappa^2$ and $\vartheta \leq 1/2$ give

$$\gamma_{i1} + \gamma_{i2} \leq 2(\kappa^2 + \vartheta) \leq 3\kappa^2. \tag{119}$$

It remains to control β_i. Add and subtract the second aligned pair of factor vectors in (117), and write $\beta_i = \beta_{i1} + \beta_{i2} + \beta_{i3} + \beta_{i4}$, where

$$\begin{aligned} \beta_{i1} &= [\Delta\mathbf{h}_i^{\mathsf{H}}\mathbf{h}_{i2}][\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}], & \beta_{i2} &= [\Delta\mathbf{h}_i^{\mathsf{H}}(\mathbf{h}_i^\star - \mathbf{h}_{i2})][\Delta\mathbf{x}_i^{\mathsf{H}}(\mathbf{x}_i^\star - \mathbf{x}_{i2})], \\ \beta_{i3} &= [\Delta\mathbf{h}_i^{\mathsf{H}}(\mathbf{h}_i^\star - \mathbf{h}_{i2})][\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}], & \beta_{i4} &= [\Delta\mathbf{h}_i^{\mathsf{H}}\mathbf{h}_{i2}][\Delta\mathbf{x}_i^{\mathsf{H}}(\mathbf{x}_i^\star - \mathbf{x}_{i2})]. \end{aligned}$$

The local condition and $\vartheta \leq 1$ imply $\|\mathbf{h}_{i2}\|_2, \|\mathbf{x}_{i2}\|_2 \leq 1 + \vartheta \leq 2$. Consequently,

$$|\beta_{i2}| + |\beta_{i3}| + |\beta_{i4}| \leq (\vartheta^2 + 4\vartheta)\|\Delta\mathbf{h}_i\|_2\|\Delta\mathbf{x}_i\|_2 \leq 5\vartheta\|\Delta\mathbf{h}_i\|_2\|\Delta\mathbf{x}_i\|_2 \leq \frac{5\vartheta}{4}\|\mathbf{u}_i\|_2^2. \tag{120}$$

The first-order stationarity condition for the alignment problem (107) at $\iota = 1$ is

$$\|\Delta\mathbf{h}_i\|_2^2 + \Delta\mathbf{h}_i^{\mathsf{H}}\mathbf{h}_{i2} = \|\Delta\mathbf{x}_i\|_2^2 + \mathbf{x}_{i2}^{\mathsf{H}}\Delta\mathbf{x}_i. \tag{121}$$

Equation (121) gives directly

$$\beta_{i1} = \left(\|\Delta\mathbf{x}_i\|_2^2 - \|\Delta\mathbf{h}_i\|_2^2 + \mathbf{x}_{i2}^{\mathsf{H}}\Delta\mathbf{x}_i\right)\left(\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}\right). \tag{122}$$

Since $\|\Delta\mathbf{x}_i\|_2^2 - \|\Delta\mathbf{h}_i\|_2^2$ is real,

$$\begin{aligned} \beta_{i1} + \beta_{i1}^* &= 2\,\mathrm{Re}\left(\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}\right)\left(\|\Delta\mathbf{x}_i\|_2^2 - \|\Delta\mathbf{h}_i\|_2^2\right) + 2\left|\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}\right|^2 \\ &\geq -2\left|\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}\right|\left|\|\Delta\mathbf{x}_i\|_2^2 - \|\Delta\mathbf{h}_i\|_2^2\right| \geq -\left|\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}\right|\|\mathbf{u}_i\|_2^2 \geq -4\vartheta\|\mathbf{u}_i\|_2^2. \end{aligned} \tag{123}$$

For the last inequality, we used $\|\Delta\mathbf{x}_i\|_2 \le 2\vartheta$ and $\|\mathbf{x}_{i2}\|_2 \le 1+\vartheta$, which imply $|\Delta\mathbf{x}_i^{\mathsf{H}}\mathbf{x}_{i2}| \le 2\vartheta(1+\vartheta) \le 4\vartheta$. Combining (120) and (123) yields

$$\beta_i + \beta_i^* \ge -\frac{13\vartheta}{2}\,\|\mathbf{u}_i\|_2^2\,. \tag{124}$$

Substituting (118), (119), and (124) into (116) gives

$$\mathrm{Re}\big(\mathbf{u}_i^{\mathsf{H}}\mathbf{P}_i\mathbf{H}_i\mathbf{u}_i\big) \ge [1-(19.5\kappa^2+1)\vartheta]\,\|\mathbf{u}_i\|_2^2 \ge \frac{1}{2}\,\|\mathbf{u}_i\|_2^2\,, \tag{125}$$

where the second inequality follows from the assumed upper bound on ϑ. Finally, both $\mathbf{P}$ and $\mathbf{H}$ are Hermitian, so

$$\mathbf{u}^{\mathsf{H}}(\mathbf{PH}+\mathbf{HP})\mathbf{u} = 2\sum_{i=1}^{n}\mathrm{Re}\big(\mathbf{u}_i^{\mathsf{H}}\mathbf{P}_i\mathbf{H}_i\mathbf{u}_i\big) \ge \sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2 = \|\mathbf{u}\|_2^2\,.$$

This proves (111). The cancellation is explicit: the ith population Hessian block contributes σ_i^2, while the sourcewise normalization contributes σ_i^{-2}. Their product is one for every source, which is precisely why the population Stage-II rate in Figure 5 is unchanged as κ varies. □