DECENTRALIZED PHASE RETRIEVAL: STATISTICAL GUARANTEES AND COMMUNICATION COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider noiseless real Gaussian measurements and use the same mix–gradient–mix update at every iteration. We couple the network average to a population gradient trajectory and control the disagreement at each gradient evaluation. This gives a finite-accuracy recovery guarantee with explicit sample, iteration, and communication costs, including unequal batches and nodes with fewer than d samples. The bound is conservative and polynomial in the dimension, initialization scale, and target accuracy. A separate local result gives exact convergence under additional regularity. Experiments show two-stage dynamics, recovery from individually underdetermined nodes, and a topology-dependent communication cost. To the best of our knowledge, this work provides the first systematic study jointly addressing statistical recovery, convergence, and communication complexity for decentralized phase retrieval over peer-to-peer networks from independent Gaussian node initializations.

1 INTRODUCTION

Phase retrieval seeks a signal from the magnitudes of its linear measurements. It arises when a detector records intensity but not phase, as in X-ray crystallography, coherent diffraction imaging, and optical imaging (Millane, 1990; Shechtman et al., 2015). Distributed acquisition provides a further motivation: the Event Horizon Telescope (EHT) combined measurements from geographically separated radio telescopes to image the black-hole shadow in M87* (Event Horizon Telescope Collaboration, 2019b; European Southern Observatory, 2019). Its imaging pipeline combines visibilities, calibration information, and closure quantities (Event Horizon Telescope Collaboration, 2019a). These applications involve complex and structured sensing operators; they motivate our setting, while the independent real Gaussian intensity model studied here serves as a tractable theoretical benchmark.

We ask whether an iterative method can recover one signal from independent random starts when measurements remain at their acquisition nodes and only neighbor communication is available. Figure 1 illustrates this setting: each node stores sensing vectors and intensities, maintains an estimate of the same signal, and communicates only with its neighbors over a fixed network. The signed projections remain implicit in the intensity measurements, and the goal is an accurate estimate at every node with one common sign. Keeping raw data local defines the computation model; the central question is how the pooled measurements and limited communication jointly determine recovery.

Centralized gradient descent can recover this model from a random start (Chen et al., 2019), but the network problem introduces additional difficulties. The initial signal correlation is small, local gradients are heterogeneous, and the quartic loss is only locally smooth, so communication errors can grow before the reference trajectory reaches its local region. Sign symmetry creates a further issue: nodewise sign-invariant error does not imply consensus. To address these challenges, we develop a neighbor-only mix–gradient–mix method with unequal sample weights and independent Gaussian starts, without a shared direction or a warm-start phase. Theorem 1 gives explicit sample, iteration, and gossip budgets for finite-accuracy recovery to one common sign, allowing nodes with fewer than d observations. Our analysis separates early signal growth from later refinement through a population reference, transfers this trajectory to the fixed-data network using a uniform fourth-moment event, and establishes an invariant region controlling both tracking error and disagreement. Proposition 4 further gives exact local recovery and consensus under additional curvature and step-size assumptions. The

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

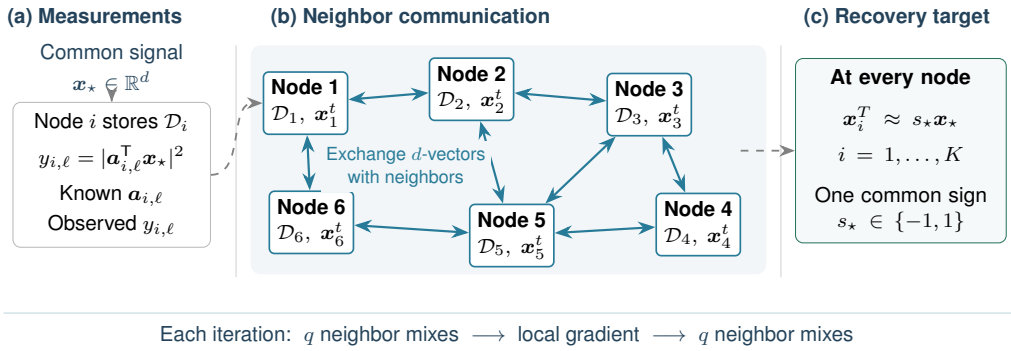


Figure 1: Phase retrieval over a peer-to-peer network. Each node stores a local intensity dataset $\mathcal{D}_i$ and starts from its own Gaussian vector. Solid edges carry estimate or post-gradient vectors, while dashed arrows indicate the measurement model and recovery target. Every iteration uses the same mix–gradient–mix update.

global sample bound is conservative, with a large polynomial dependence on the problem parameters and target accuracy; a near-linear guarantee for independent distributed starts remains open.

1.1 Related work

Centralized phase retrieval: classical methods and convex relaxations. Classical methods alternate between enforcing measured magnitudes and constraints in the signal domain. The Gerchberg–Saxton iteration and Fienup's error-reduction and input–output algorithms established this approach for optical imaging (Fienup, 1982); their practical success motivated later statistical analyses under random sensing models. Convex formulations provide a different route to provable recovery. PhaseLift lifts the signal to a positive semidefinite rank-one matrix, making intensity measurements linear in the lifted variable, and relaxes recovery to semidefinite optimization (Candès et al., 2013). Its original analysis established exact recovery and noise stability with $O(d \log d)$ random measurements; Candès & Li (2014) subsequently reduced the sufficient measurement count to $O(d)$. PhaseCut instead relaxes a quadratic optimization problem over the unknown measurement phases (Waldspurger et al., 2015). These semidefinite approaches introduce matrix variables, which increase storage and computational costs. PhaseMax avoids lifting by optimizing a linear objective over convex magnitude constraints in the original signal space; its recovery guarantee depends on the quality of an anchor direction and the number of measurements (Goldstein & Studer, 2017).

Spectral initialization and nonconvex refinement. Scalable methods typically construct an informative initial estimate and then refine it in the signal domain. Netrapalli et al. (2013) proved geometric convergence of a spectrally initialized, resampled alternating-minimization scheme under Gaussian sensing, with sample complexity near the dimension up to logarithmic factors. Wirtinger flow combines spectral initialization with gradient descent on the intensity-based quartic loss and gives a recovery guarantee from $O(d \log d)$ Gaussian measurements (Candès et al., 2015). Truncated Wirtinger flow modifies the initialization and suppresses overly influential gradient terms, obtaining linear sample complexity and improved computational guarantees (Chen & Candès, 2015). Amplitude-based methods change the objective: truncated amplitude flow uses orthogonality-promoting initialization and truncated generalized-gradient updates (Wang et al., 2018), while reshaped Wirtinger flow achieves linear convergence without gradient truncation (Zhang & Liang, 2016). Both admit recovery guarantees with $O(d)$ Gaussian measurements under their respective initialization conditions. These results show how the objective, initialization, and update rule affect statistical and computational efficiency; they do not by themselves justify applying an unmodified quartic gradient iteration from independent random node states.

Global geometry and implicit regularization. A complementary line of work studies why nonconvex optimization succeeds. Under complex Gaussian sensing with sufficiently many measurements, Sun et al. (2018) showed that the quartic loss has no spurious local minima and has negative curvature at saddle points, and analyzed a trust-region method that does not require a special initialization.

Such geometric results distinguish globally optimal solutions from other stationary points, but do not alone establish the iteration complexity of vanilla gradient descent. For gradient descent started from a spectral estimate, Ma et al. (2018) showed that the iterates remain in a region with favorable curvature and controlled correlations with the sensing vectors. Their leave-one-out analysis explains how implicit regularization permits faster convergence without explicit trimming or projection, while accounting for the dependence between iterates and a reused dataset.

Recovery from random or arbitrary initialization. The closest centralized benchmark is Chen et al. (2019), who proved that randomly initialized gradient descent on the real Gaussian quartic loss reaches relative accuracy ε in $O(\log d + \log(1/\varepsilon))$ iterations with sample complexity linear in d up to logarithmic factors. The proof tracks an initial signal-amplification stage followed by local refinement, and uses leave-one-out arguments without sample splitting. For the nonsmooth amplitude loss, Tan & Vershynin (2023) established convergence of online stochastic gradient descent, equivalently a randomized Kaczmarz scheme, from arbitrary initialization; that analysis uses fresh measurements and therefore differs from repeated optimization over one fixed dataset. More recently, Li et al. (2025) established a random-initialization guarantee for reshaped Wirtinger flow with nearly linear sample complexity and logarithmic iteration complexity. These centralized results motivate our analysis, but they track a single estimate with access to pooled or freshly sampled data. Our network setting requires simultaneous control of a weak initial average, heterogeneous local gradients, finite gossip, and agreement on one common sign. We establish explicit sufficient statistical and communication budgets for this setting, while our uniform fourth-moment argument gives a substantially more conservative sample bound than the centralized trajectory analyses.

Distributed optimization and estimation. Distributed optimization combines local objective updates with neighbor communication to minimize a sum of nodewise losses. Consensus-based subgradient methods establish convergence for convex objectives (Nedić & Ozdaglar, 2009); EXTRA enables exact convergence with a fixed step size in smooth convex problems (Shi et al., 2015), while NEXT combines successive convex approximation and gradient tracking for nonconvex optimization (Di Lorenzo & Scutari, 2016). In distributed estimation, diffusion strategies let nodes learn cooperatively from streaming observations, with analyses of stability and mean-square performance (Sayed, 2014). For structured estimation, DeEPCA uses subspace tracking and consensus to achieve linear convergence in decentralized principal component analysis (Ye & Zhang, 2021), and Maros & Scutari (2023) establish statistical and convergence guarantees for decentralized matrix sensing from random initialization under an in-network RIP. These approaches connect communication with optimization and estimation accuracy; phase retrieval additionally requires handling the Gaussian intensity geometry and the global sign ambiguity.

Distributed phase retrieval. Several methods have extended phase retrieval to distributed architectures. The primal–dual distributed Wirtinger flow (DWF) of Zhao et al. (2018) converges to approximate KKT points at a sublinear rate; this optimization guarantee does not establish finite-sample recovery of the underlying signal. Other approaches include asynchronous ADMM over connected networks (Mafakheri et al., 2024), decentralized expectation-consistent inference with state-evolution analysis (Wang et al., 2020), and server-based multi-view recovery with outliers (Yang, 2025). Despite these advances, a unified understanding of the statistical recovery properties and linear convergence rates of peer-to-peer phase retrieval remains incomplete, particularly for independent random node initializations and fixed local datasets.

2 Preliminaries and problem formulation

2.1 Measurements and the global objective

We consider a network of K nodes that seek the same unknown vector $\boldsymbol{x}_\star \in \mathbb{R}^d$. Its norm is denoted by r. Node i holds m_i measurements, and M denotes the total number of measurements in the network. Each sensing vector is standard Gaussian. The detector records the squared projection onto the signal, so the sign of that projection is lost:

$$y_{i,\ell} = (\boldsymbol{a}_{i,\ell}^\top \boldsymbol{x}_\star)^2, \quad \boldsymbol{a}_{i,\ell} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad M = \sum_{i=1}^{K} m_i. \tag{1}$$

The sample counts and their assignment to nodes are fixed independently of the draws. Because the observations are unchanged when $\boldsymbol{x}_\star$ is replaced by $-\boldsymbol{x}_\star$, recovery is possible only up to one global sign.

To estimate the signal, each node measures the mismatch between predicted and observed intensities. We write this local quartic loss as f_i. The pooled loss F weights each f_i by its share of the total data, so F is exactly the empirical loss that would be formed if all M measurements were centralized:

$$f_i(\boldsymbol{x}) = \frac{1}{4m_i}\sum_{\ell=1}^{m_i}[(\boldsymbol{a}_{i,\ell}^\top \boldsymbol{x})^2 - y_{i,\ell}]^2, \qquad F(\boldsymbol{x}) = \sum_i \frac{m_i}{M} f_i(\boldsymbol{x}). \tag{2}$$

Consensus mixing gives every node the same weight, whereas the pooled loss weights nodes in proportion to m_i. We therefore rescale the local objective before differentiating it. The resulting vector $\boldsymbol{g}_i$ is the gradient actually used by node i:

$$h_i(\boldsymbol{x}) = \frac{Km_i}{M} f_i(\boldsymbol{x}), \qquad \boldsymbol{g}_i(\boldsymbol{x}) = \nabla h_i(\boldsymbol{x}) = \frac{K}{M}\sum_{\ell=1}^{m_i}[(\boldsymbol{a}_{i,\ell}^\top \boldsymbol{x})^2 - y_{i,\ell}](\boldsymbol{a}_{i,\ell}^\top \boldsymbol{x})\boldsymbol{a}_{i,\ell}. \tag{3}$$

This choice ensures that the uniform average of the scaled local objectives equals F, even when the batch sizes differ. Without it, the network would optimize an unweighted average of local losses. In the noiseless model, both signed solutions make every local gradient vanish.

2.2 Network and independent initialization

Communication takes place on a fixed connected graph. A matrix $\boldsymbol{W}$ specifies one gossip round: W_{ij} is the weight that node i assigns to the vector received from node j. We assume that $\boldsymbol{W}$ is symmetric, nonnegative, doubly stochastic, has a positive diagonal, and uses only graph edges. The matrix $\boldsymbol{J}$ replaces every node state by their exact average, while ρ measures how far one gossip round is from this ideal operation:

$$\boldsymbol{J} = \mathbf{1}\mathbf{1}^\top/K, \qquad \rho = \|\boldsymbol{W} - \boldsymbol{J}\|_2 < 1. \tag{4}$$

After q gossip rounds, disagreement contracts by at most ρ^q. The nodes know K, M, and a certified upper bound on ρ. The special case $\boldsymbol{W} = \boldsymbol{J}$ represents exact averaging and needs only one round.

The algorithm starts without a shared direction. We assume only public lower and upper bounds r_- and r_+ on the unknown norm r. The scalar τ controls the size of the random start, and $\sigma_0 = \tau r_-$ is its root-mean-square norm. Each node independently draws

$$\boldsymbol{v}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \qquad \boldsymbol{x}_i^0 = \frac{\sigma_0}{\sqrt{d}}\,\boldsymbol{v}_i. \tag{5}$$

Thus the nodes may initially point in unrelated directions and may even have opposite correlations with the truth. The algorithm neither normalizes these vectors nor coordinates their random seeds. Their average is smaller by a factor of $K^{-1/2}$; Appendix B records its distribution and the associated probability event. The remaining public calibration and metadata assumptions are also collected in Appendix A.

We use two errors because accurate nodewise reconstruction does not by itself imply that the nodes agree on the same sign. Stack the node estimates as the rows of $\boldsymbol{X}^t$. Their average is $\bar{\boldsymbol{x}}^t$, and Δ_t is the root-mean-square distance from consensus. The common-sign error E_{com} then measures reconstruction after choosing one sign for the entire network:

$$\bar{\boldsymbol{x}}^t = K^{-1}\sum_i \boldsymbol{x}_i^t, \qquad \Delta_t = K^{-1/2}\left\|(\boldsymbol{I}_K - \boldsymbol{J})\boldsymbol{X}^t\right\|_F,$$
$$E_{\mathrm{com}}(t) = \min_{s\in\{-1,1\}}\left(\frac{1}{K}\sum_i \left\|\boldsymbol{x}_i^t - s\boldsymbol{x}_\star\right\|^2\right)^{1/2}. \tag{6}$$

We also report $E_{\max}(t) = \max_i \min_s \|\boldsymbol{x}_i^t - s\boldsymbol{x}_\star\|$. Unlike $E_{\max}$, E_{com} uses one sign for all nodes. Thus two nodes at $\boldsymbol{x}_\star$ and $-\boldsymbol{x}_\star$ have zero nodewise sign-invariant error but nonzero disagreement. The theorem controls the stronger common-sign error.

Although phase retrieval can be lifted to a rank-one matrix problem, Gaussian intensity measurements have a trace-corrected fourth-moment geometry rather than the isotropic geometry of standard matrix sensing. Appendix D gives the identities and a finite-sample obstruction. This distinction explains why we prove a pooled fourth-moment bound instead of invoking an in-network matrix-sensing RIP.

Algorithm 1 GOSSIP-PR: independent Gaussian starts and mix–gradient–mix

Require: Local data; $M, K, \boldsymbol{W}, \bar{\rho}, r_-, r_+$; $\tau \in (0, 1/4]$; $\varepsilon, p, \zeta \in (0, 1)$.

1: Set $\sigma_0 = \tau r_-$ and $\gamma_0 = \sigma_0/\sqrt{K}$. Draw $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ independently at each node; set $\boldsymbol{x}_i^0 = (\tau r_-/\sqrt{d})\boldsymbol{v}_i$.
2: Configure the step size η, horizon T, and mixing count q as in Appendix A.
3: **for** $t = 0, \dots, T-1$ **do**
4: Set $\boldsymbol{u}_i = \boldsymbol{x}_i^t$. For q rounds, exchange $\boldsymbol{u}_i$ and simultaneously replace $\boldsymbol{u}_i \leftarrow \sum_j W_{ij}\boldsymbol{u}_j$.
5: Set $\boldsymbol{y}_i^t = \boldsymbol{u}_i$ and compute $\boldsymbol{u}_i = \boldsymbol{y}_i^t - \eta \boldsymbol{g}_i(\boldsymbol{y}_i^t)$ using the fixed local dataset.
6: For q rounds, exchange $\boldsymbol{u}_i$ and simultaneously replace $\boldsymbol{u}_i \leftarrow \sum_j W_{ij}\boldsymbol{u}_j$.
7: Set $\boldsymbol{x}_i^{t+1} = \boldsymbol{u}_i$.
8: **end for**
9: Return $\boldsymbol{x}_i^T$. The horizon is fixed from the public inputs and does not require the unknown signal.

3 ALGORITHM

One iteration has three operations: nodes first average their current estimates for q gossip rounds, then take local gradient steps, and finally average the updated vectors for another q rounds. To write these operations compactly, let $\boldsymbol{P} = \boldsymbol{W}^q$ be the q-round mixing matrix. The rows of $\boldsymbol{Y}^t$ are the vectors at which gradients are evaluated, and the rows of $\boldsymbol{G}(\boldsymbol{Y}^t)$ are the corresponding scaled local gradients. The update is

$$\boldsymbol{Y}^t = \boldsymbol{P}\boldsymbol{X}^t, \qquad \boldsymbol{X}^{t+1} = \boldsymbol{P}[\boldsymbol{Y}^t - \eta \boldsymbol{G}(\boldsymbol{Y}^t)], \qquad \boldsymbol{G}(\boldsymbol{Y})_{i,:} = \boldsymbol{g}_i(\boldsymbol{y}_i)^\top. \tag{7}$$

The first mixing block makes the gradient-evaluation points more alike. This matters because otherwise different nodes may evaluate a nonlinear gradient at widely separated vectors. The local step then uses the fixed data stored at each node. Since those datasets are different, the local steps introduce new disagreement; the second block suppresses it before the next iteration. Each message is a d-vector, while raw measurements and lifted matrices remain at their owners. The first block is part of the first iteration, so the initialization itself remains genuinely independent across nodes.

The theorem uses a fixed step size, a fixed iteration horizon, and enough mixing to keep the network perturbation below the desired reconstruction accuracy. Appendix A gives the complete expressions and explains the role of each constant. They are sufficient proof parameters rather than practical tuning recommendations; the experiments use the simpler settings reported in Appendix I.

4 MAIN RESULTS AND CONVERGENCE ANALYSIS

4.1 RECOVERY GUARANTEE AND ITS INTERPRETATION

The main theorem answers three questions at once. It shows that independent nodewise starts can select one common sign, that all nodes reach a prescribed relative accuracy, and that finite gossip is sufficient to keep them close to consensus. Fix a target accuracy ε and failure budgets p and ζ. The former controls an unusually weak initial correlation with the signal; the latter controls the fixed-data concentration events. The complete parameter schedule and sample condition are stated in Appendix A.

Theorem 1 (Recovery from independent Gaussian starts)**.** *Under the measurement, network, and calibration assumptions above, run Algorithm 1 with the schedule in Appendix A. If the total sample size satisfies* (16)*, then, with probability at least* $1 - p - \zeta - 2e^{-d/4}$*, there is one sign* $s_\star \in \{-1, 1\}$ *shared by all nodes such that*

$$\max_i \left\| \boldsymbol{x}_i^T - s_\star \boldsymbol{x}_\star \right\| \le \varepsilon r, \qquad E_{\mathrm{com}}(T) \le \varepsilon r, \qquad \Delta_T \le D. \tag{8}$$

All gradient-evaluation vectors $\boldsymbol{y}_i^t$*,* $0 \le t < T$*, and all output states* $\boldsymbol{x}_i^t$*,* $1 \le t \le T$*, have norm at most* R*. The raw independent states at* $t = 0$ *may lie outside this ball and may have different correlation signs; the proof only requires* $m_i \ge 1$*.*

The first bound in (8) is the strongest: every node is close to the same signed signal. It immediately implies the common-sign root-mean-square bound. The last bound separately quantifies

consensus, which is useful even when reconstruction error is already small. No node needs enough local measurements to identify the signal on its own; recovery is driven by the pooled data and communication.

Remark 2 (Scope of the recovery guarantee). The theorem is a finite-accuracy result for noiseless real Gaussian measurements on a synchronous graph. Its sufficient sample size is polynomial and much larger than the near-linear behavior seen in centralized theory and in our experiments; Appendix F gives the explicit order. The theorem permits individually underdetermined nodes, but it does not certify every balanced instance in which all local batches are underdetermined. Sharper sample rates, arbitrary accuracy at a fixed sample size, and noisy or complex measurements remain open.

Each of the T iterations evaluates one full local gradient at every node and uses two blocks of q gossip rounds. Across the network, this amounts to MT sample-gradient evaluations and $2qT$ communication rounds. Because every message is a d-vector, a sparse graph can send fewer words than a complete graph even when it needs more rounds. Appendix F gives the exact word counts and the dependence of q on the spectral gap.

The complete-mixing limit. The role of the two mixing blocks is clearest under exact averaging. When $\boldsymbol{W} = \boldsymbol{J}$, every node evaluates its gradient at the current network average, and the second block averages the correctly weighted local gradients. Thus all rows of the next state equal one centralized gradient step:

$$\boldsymbol{X}^{t+1} = \mathbf{1}[\bar{\boldsymbol{x}}^t - \eta \nabla F(\bar{\boldsymbol{x}}^t)]^\top.$$

The method then reproduces centralized empirical gradient descent from the average of the independent initial states. On a sparse graph, finite mixing introduces the two perturbations analyzed below. The population reference is a proof-level comparison trajectory, distinct from the empirical centralized path executed by the nodes.

4.2 Why recovery occurs

The proof follows an ideal trajectory and then shows that the distributed iterates remain close to it. Let F_{pop} be the expected loss under a fresh Gaussian measurement. Starting from the average of the independent node states, the reference vector $\boldsymbol{z}^t$ follows gradient descent on this population loss:

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta \nabla F_{\mathrm{pop}}(\boldsymbol{z}^t), \qquad \boldsymbol{z}^0 = \bar{\boldsymbol{x}}^0. \tag{9}$$

To see what this reference does, decompose it into a component along the true signal and an orthogonal component. The unit vector $\boldsymbol{e} = \boldsymbol{x}_\star / r$ gives the signal direction; α_t is the magnitude of the projection onto this direction, and β_t is the norm of the remaining component. Their updates are

$$\alpha_{t+1} = [1 + 3\eta(r^2 - \alpha_t^2 - \beta_t^2)]\alpha_t, \quad \beta_{t+1} = [1 + \eta(r^2 - 3\alpha_t^2 - 3\beta_t^2)]\beta_t. \tag{10}$$

While both components are small, the multiplicative factor for α_t is larger than that for β_t. The initially weak signal component therefore becomes dominant. Once it reaches a constant fraction of r, the dynamics enter a refinement stage and contract toward one signed solution. Appendix C proves these claims and shows that the reference reaches half of the final error budget by time T.

The actual algorithm reuses one finite dataset, so its gradients do not equal the population gradient exactly. Appendix D controls this difference uniformly over every vector in the region visited by the proof. Uniformity is important: the iterates depend on the same measurements, so a bound for one fixed, independent vector would not suffice. The argument uses one pooled fourth-moment event, rather than a separate RIP condition at every node.

Finite communication creates a second error because nodes evaluate gradients at different vectors. The first mixing block reduces this discrepancy before gradient evaluation, and the second contracts the disagreement introduced by the local steps. Appendix F derives coupled recursions for the distance between the network average and $\boldsymbol{z}^t$, and for the distance of individual nodes from that average. The parameter schedule keeps both quantities inside a common invariant region.

Remark 3 (Role of the empirical control). The pooled event removes any need for a nodewise RIP or a minimum local sample size. Its uniform fourth-order tensor norm is much stronger than the trajectory-specific control that recovery actually needs, however, and this strength produces the conservative polynomial sample bound. The result is therefore a sufficient transfer principle, not an optimal sampling theorem.

At the final iteration, every node's error is split into three interpretable pieces: the error of the ideal population trajectory, the error made while tracking that trajectory with the network average, and the remaining node-to-average disagreement. If a_T denotes the second term, the triangle inequality gives

$$\max_i \left\|\boldsymbol{x}_i^T - s_\star \boldsymbol{x}_\star\right\| \le \underbrace{\left\|\boldsymbol{z}^T - s_\star \boldsymbol{x}_\star\right\|}_{\text{population recovery}} + \underbrace{a_T}_{\text{mean tracking}} + \underbrace{\sqrt{K}\Delta_T}_{\text{node disagreement}} .$$

The first term is controlled by the two-stage dynamics. Uniform concentration and sufficient mixing allocate the remaining half of the error budget to the other two terms. This decomposition is the organizing idea of the proof; all auxiliary constants and recursions are deferred to the appendix.

4.3 Exact local convergence under explicit regularity

The global theorem stops at a prescribed finite accuracy. A separate local result explains what happens after the iterates enter a regular neighborhood of one signed solution. Fix such a solution $\boldsymbol{x}^\dagger$ and a ball of radius b around it. The assumptions below say that the pooled loss has curvature between μ and L throughout this ball, every local gradient is L-Lipschitz there, and all local gradients vanish at the same noiseless solution. Appendix G specifies how much mixing is sufficient.

Proposition 4 (Local invariant neighborhood). *Fix $\boldsymbol{x}^\dagger = s\boldsymbol{x}_\star$. Suppose the regularity conditions above hold, use a step $0 < \eta_{\mathrm{loc}} \le 1/L$, and choose q according to* (35)*. Let V_t combine the error of the network average and the node disagreement as defined in Appendix G. If $V_0 \le b$, then every iterate remains in the ball and*

$$\max_i \left\|\boldsymbol{x}_i^t - \boldsymbol{x}^\dagger\right\| \le V_t \le (1 - \eta_{\mathrm{loc}}\mu/2)^t V_0 \quad (t \ge 0). \tag{11}$$

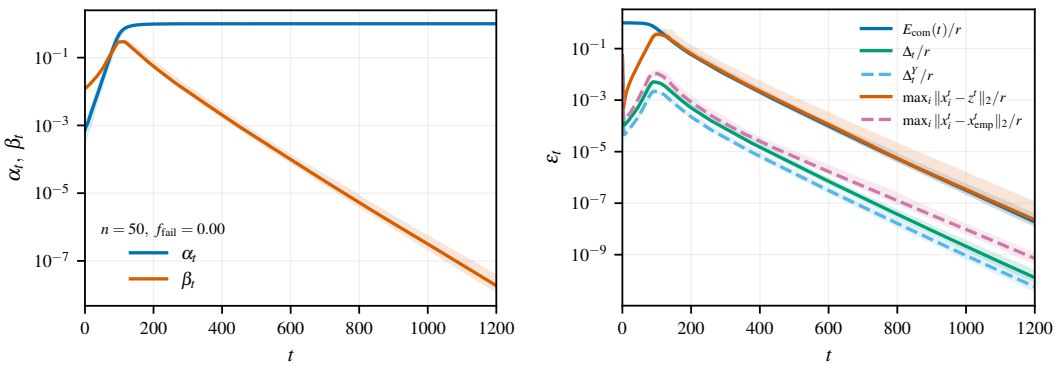
Remark 5 (Local versus global convergence). Proposition 4 is conditional: Appendix G proves that a suitable neighborhood exists almost surely for Gaussian data when $M \ge d$, but its radius and curvature depend on the realized problem. The result therefore complements rather than extends the global theorem. A global exact-rate theorem would need a computable rule certifying entry into this neighborhood. With noise, local gradients generally do not share a common zero, so a separate stability analysis is required.

5 Numerical experiments

Setup and evaluation. The experiments examine two-stage dynamics, recovery from locally insufficient batches, and the effect of topology on communication cost. Each trial uses a unit-norm signal and M independent standard Gaussian sensing vectors, with noiseless intensities fixed throughout the run. Unless otherwise stated, nodes start independently with $\tau = 0.1$ and public norm bounds $[r_-, r_+] = [1/2, 2]$. The recovery study uses $d \in \{64, 128, 256, 512\}$, $K \in \{4, 16, 64\}$, and $M/d \in \{2, 3, 4, 6, 8, 12, 16, 24, 32\}$, with $q = 16$ gossip rounds per block. Graphs are connected random 4-regular graphs, except that $K = 4$ uses the complete graph. A disjoint set of 20 pilot seeds selects the step size $\eta = 0.02$ for GOSSIP-PR.

A trial succeeds when $\min_{s\in\{\pm1\}} \max_i \|\boldsymbol{x}_i - s\boldsymbol{x}_\star\| \le 10^{-3}$ and $\Delta_t \le 10^{-4}$ within 1200 pooled full-data passes. Each pass evaluates every measurement gradient once across the network. Recovery and cooperation cells use 100 evaluation seeds, and trajectory and topology cells use 50. Failed, divergent, and censored trials remain in the success probabilities. We report Wilson 95% intervals for success fractions and medians with interquartile ranges for trajectories. Paired methods share the truth, data, partition, and initialization draws. Comparators include centralized and independent local descent with random or spectral starts, DWF, gradient tracking, and decentralized spectral initialization followed by GOSSIP-PR refinement; spectral preprocessing is included in communication costs. Appendix I gives the seed streams, comparator updates, and remaining ablations.

Two-stage dynamics. We first ask whether the population mechanism remains visible in the distributed finite-data algorithm. In a representative configuration with dimension 128, oversampling ratio 12, 16 nodes, and 16 gossip rounds per block, all 50 trials succeed. Figure 2a shows the initially weak signal component becoming dominant while the orthogonal component decays. The median iterate enters the local basin after 174 pooled passes. Figure 2b then shows geometric refinement: the final median common-sign error is 1.87×10^{-8} and the median disagreement is 1.29×10^{-10}. The centralized and population curves are offline references used to interpret the distributed path. Appendix I.5 shows a difficult configuration in which this transition fails.



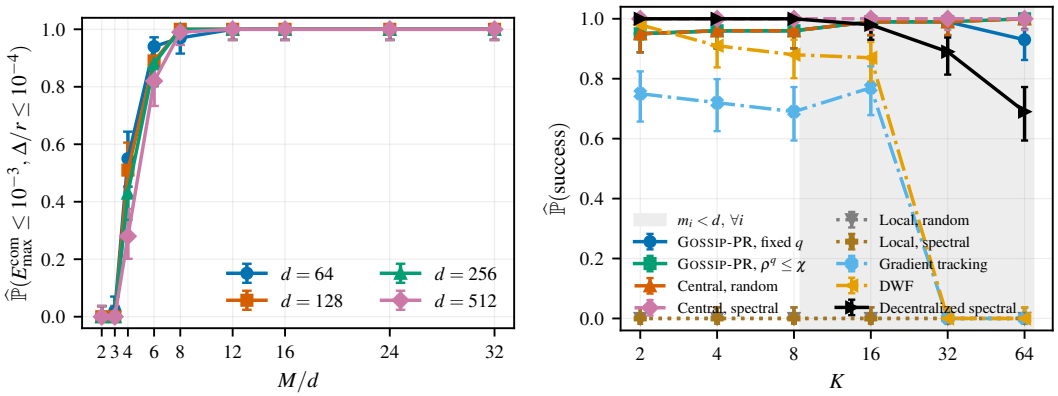
(a) Signal and orthogonal components.

(b) Network and reference errors.

Figure 2: Two-stage dynamics; medians and interquartile ranges over 50 seeds.

Recovery and cooperation. We next ask whether recovery is governed by the pooled sample count rather than by the size of each local batch. Figure 3a varies the total oversampling ratio with 16 nodes. Across dimensions 64 through 512, the success probability rises from 0.28–0.55 at ratio 4 to 0.82–0.94 at ratio 6 and at least 0.97 at ratio 8. This transition occurs far below the conservative sufficient condition in the theorem.

Figure 3b isolates the effect of cooperation by fixing 1024 samples in dimension 128 and repartitioning the same pooled data over more nodes. When there are 16, 32, or 64 nodes, every local batch contains fewer samples than the ambient dimension. Nevertheless, GOSSIP-PR with 16 gossip rounds succeeds in 0.93–0.99 of trials, whereas methods using only local data never succeed. Increasing the mixing count until the residual contraction is at most 10^{-2} matches the centralized random-start success rate, including probability 1.00 at 64 nodes. The experiment supports the value of pooling through communication; it does not sharpen the theorem's sample condition.



(a) Recovery transition at $K = 16$.

(b) Fixed $(d, M) = (128, 1024)$.

Figure 3: Recovery probability with Wilson 95% intervals.

The same cooperation effect persists at a larger dimension. With $(d, M) = (256, 2048)$, Figure 4 shows that fixed-q GOSSIP-PR succeeds on 0.88 of trials at $K = 64$, where each node has only $m_i/d = 1/8$ of the ambient dimension in samples. The attenuation-target variant and centralized random-start descent reach 0.98, while independent local methods never recover. This higher-dimensional check supports the pooled-data interpretation without changing the theorem's conservative sample requirement.

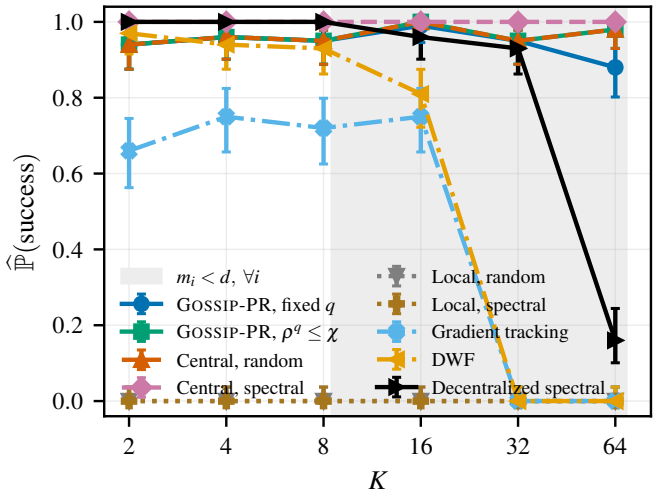


Figure 4: Fixed pooled data at $(d, M) = (256, 2048)$.

Topology and communication. Finally, we ask whether faster mixing always means cheaper communication. We compare five graph families with 16 nodes, dimension 128, and 1024 pooled samples. Complete graphs use $\boldsymbol{W} = \boldsymbol{J}$, and sparse graphs use symmetric Metropolis weights. Writing $T_{10^{-3}}$ for the first successful iteration, we count $T_{10^{-3}}$ pooled passes, $2qT_{10^{-3}}$ gossip rounds, and $4|\mathcal{E}|dqT_{10^{-3}}$ transmitted scalar words. All recover reliably, but their ranking depends on whether cost is measured in rounds or transmitted words. With one gossip round per block, the median cost of reaching the target is 3.59×10^7 words on the complete graph and 5.24×10^6 on the cycle. The complete graph mixes faster, but each round uses many more edges. Thus a sparse topology can reduce total traffic despite requiring more rounds. Appendix I.5 reports the separate gradient-pass, gossip-round, and word counts for 16 and 64 nodes.

6 Conclusion and Limitations

In this paper, within the noiseless real Gaussian model, we have analyzed the repeated mix–gradient–mix update and have established finite-accuracy recovery and consensus guarantees. Our experiments have demonstrated two-stage dynamics, pooled recovery from underdetermined local data, and lower transmitted volume on sparse graphs. We have also proved exact local convergence under additional regularity assumptions. To the best of our knowledge, we have provided the first systematic study jointly addressing statistical recovery, convergence, and communication complexity for decentralized phase retrieval over peer-to-peer networks from independent Gaussian node initializations.

Our analysis has been limited to the calibrated, noiseless, real Gaussian model on a synchronous network. The pooled fourth-moment argument has yielded a conservative polynomial sample requirement and has not explained the sharper empirical transition. The local result has established exact convergence only within a regular neighborhood, and we have not provided a computable rule for certifying entry into that neighborhood. These limitations have highlighted directions for future work: sharper sample bounds through trajectory-specific concentration or leave-one-out arguments, a certified basin-entry mechanism for global exact rates, and extensions to noisy, complex-valued, and structured imaging measurements, including EHT-specific models.

AI USE STATEMENT

Generative AI (OpenAI Codex) assisted with literature review, mathematical and proof development, experiment design and execution, figure generation, writing, and LaTeX compilation. The authors retain responsibility for the mathematics, implementation, empirical interpretation, novelty assessment, and this disclosure.

REFERENCES

Emmanuel J. Candès and Xiaodong Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5): 1017–1026, 2014. doi: 10.1007/s10208-013-9162-z.

Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. doi: 10.1002/cpa.21432.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. doi: 10.1109/TIT.2015.2399924.

Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

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.

Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2016.

European Southern Observatory. Astronomers capture first image of a black hole. ESO press release eso1907, April 10, 2019.

Event Horizon Telescope Collaboration. First M87 Event Horizon Telescope results. IV. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019a. doi: 10.3847/2041-8213/ab0e85.

Event Horizon Telescope Collaboration. First M87 Event Horizon Telescope results. I. the shadow of the supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L1, 2019b. doi: 10.3847/2041-8213/ab0ec7.

James R. Fienup. Phase retrieval algorithms: A comparison. *Applied Optics*, 21(15):2758–2769, 1982. doi: 10.1364/AO.21.002758.

Tom Goldstein and Christoph Studer. Convex phase retrieval without lifting via PhaseMax. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1273–1281, 2017.

Linbin Li, Haiyang Peng, Yong Xia, and Meng Huang. Convergence analysis of reshaped Wirtinger flow with random initialization for phase retrieval. arXiv preprint arXiv:2507.15684, 2025.

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, 2018.

Behnam Mafakheri, Jonathan H. Manton, and Iman Shames. A decentralised asynchronous optimisation algorithm with an application to phase retrieval. In *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2024. doi: 10.1109/SAM60225.2024.10636375.

Marie Maros and Gesualdo Scutari. Decentralized matrix sensing: Statistical guarantees and fast convergence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40154–40166, 2023.

Rick P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, 1990. doi: 10.1364/JOSAA.7.000394.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. doi: 10.1109/TAC.2008.2009515.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Ali H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7(4–5):311–801, 2014. doi: 10.1561/2200000051.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. doi: 10.1109/MSP.2014.2352673.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18:1131–1198, 2018.

Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *Journal of Machine Learning Research*, 24(58): 1–47, 2023.

Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Mathematical Programming*, 149(1–2):47–81, 2015.

Chang-Jen Wang, Chao-Kai Wen, Shang-Ho Tsai, and Shi Jin. Decentralized expectation consistent signal recovery for phase retrieval. *IEEE Transactions on Signal Processing*, 68, 2020. doi: 10.1109/TSP.2020.2974711.

Gang Wang, Georgios B. Giannakis, and Yonina C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2018. doi: 10.1109/TIT.2017.2756858.

Ming-Hsun Yang. Robust distributed phase retrieval for multi-view compressive networked sensing with outliers. arXiv preprint arXiv:2505.24651, 2025.

Haishan Ye and Tong Zhang. DeEPCA: Decentralized exact PCA with linear convergence rate. *Journal of Machine Learning Research*, 22(238):1–27, 2021.

Huishuai Zhang and Yingbin Liang. Reshaped Wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ziping Zhao, Songtao Lu, Mingyi Hong, and Daniel P. Palomar. Distributed optimization for generalized phase retrieval over networks. In *52nd Asilomar Conference on Signals, Systems, and Computers*, 2018. doi: 10.1109/ACSSC.2018.8645496.

A Parameter schedule for the recovery theorem

This section collects the proof parameters used by Theorem 1. Keeping them here separates the theorem's statistical and communication content from the constants needed to close the invariant-region argument.

Assume $K \geq 2$, $d \geq 3$, and $0 < r_- \leq r \leq r_+$, with $r_+/r_- \leq \kappa_0$ for a fixed calibration ratio. Choose an initialization fraction $0 < \tau \leq 1/4$, a target error $0 < \varepsilon < 1$, and probability budgets

$p, \zeta \in (0,1)$. The nodewise initialization scale is $\sigma_0 = \tau r_-$, and the scale of its network average is $\gamma_0 = \sigma_0/\sqrt{K}$.

The step size is $\eta = c_0/r_+^2$ for any fixed $0 < c_0 \le 1/12$. To give the weak initial signal component enough time to grow and then refine, define $s_- = \eta r_-^2$ and use

$$H_0 = \left\lceil \frac{1}{s_-} \log \frac{16 r_+ \sqrt{d}}{\gamma_0 p \varepsilon} \right\rceil, \qquad T = 4H_0 + 1. \tag{12}$$

Here p controls the small-ball event for the initial correlation, while the factor four allocates time to alignment, growth, and refinement.

The next constants describe the ball containing every gradient-evaluation point and the empirical accuracy needed inside it:

$$\begin{aligned} &R = 2r_+, && L_{\mathrm{pop}} = 39 r_+^2, && A_{\mathrm{pop}} = 1 + \eta L_{\mathrm{pop}}, \\ &S_R = R^3 + r_+^2 R = 10 r_+^3, && h = \varepsilon r_-/2, && \nu = \frac{h L_{\mathrm{pop}}}{4 S_R A_{\mathrm{pop}}^T}. \end{aligned} \tag{13}$$

Specifically, L_{pop} is a population-gradient Lipschitz bound, A_{pop} is the corresponding one-step amplification bound, S_R converts fourth-moment error into gradient error, h is the tracking-error allowance, and ν is the required tensor accuracy.

The local derivative bounds depend on a high-probability upper bound B for the sensing-vector norms. The constants L and G bound the local Hessians and gradients, R_0 bounds the initial network state, D is the allowed disagreement, and C_D is its one-step envelope:

$$\begin{aligned} &B = \sqrt{d} + \sqrt{2\log(2M/\zeta)}, \quad L = 13 K B^4 r_+^2, && G = 10 K B^4 r_+^3, \\ &R_0 = 2\gamma_0 \sqrt{K}, && D = \min\left\{ \frac{h}{2\sqrt{K}}, \frac{h L_{\mathrm{pop}}}{4 L A_{\mathrm{pop}}^T} \right\}, \quad C_D = (1+\eta L) D + \eta G. \end{aligned} \tag{14}$$

If $0 < \rho \le \bar{\rho} < 1$, choose enough gossip rounds to contract the initial and newly injected disagreement:

$$q = \max\left\{ 1, \left\lceil \frac{\log \max\{2, R_0/D, C_D/D\}}{-\log \bar{\rho}} \right\rceil \right\}; \qquad q = 1 \text{ if } \boldsymbol{W} = \boldsymbol{J}. \tag{15}$$

Finally, the pooled fourth-moment event used in the proof holds with the required probability when

$$M \ge \frac{2d(d+2)(d+4)(d+6)}{\zeta \nu^2}. \tag{16}$$

All quantities are functions of public inputs. The total count M may be supplied as metadata; otherwise tagged counts can be flooded in at least $\mathrm{diam}(\mathcal{G})$ rounds using $O(K)$ scalar words per directed edge. These sufficient values are designed for the proof and are not practical tuning recommendations.

B Independent initialization and probability events

The initial states are those in (5), and all draws are independent of the measurements. Their average serves only as a mathematical reference for the proof. Because the nodewise starts are independent Gaussians, their average is also Gaussian, with root-mean-square scale γ_0:

$$\boldsymbol{z}^0 = \bar{\boldsymbol{x}}^0 = K^{-1} \sum_i \boldsymbol{x}_i^0 \sim \mathcal{N}(\boldsymbol{0}, \gamma_0^2 \boldsymbol{I}_d/d). \tag{17}$$

Lemma 6 (Independent Gaussian initialization)**.** *For a fixed unit vector $\boldsymbol{e}$, with probability at least $1 - p - 2e^{-d/4}$,*

$$|\langle \bar{\boldsymbol{x}}^0, \boldsymbol{e} \rangle| \ge \frac{\gamma_0 p}{\sqrt{d}}, \qquad \|\bar{\boldsymbol{x}}^0\| \le 2\gamma_0 \le r/2, \qquad \frac{\|\boldsymbol{X}^0\|_F}{\sqrt{K}} \le 2\gamma_0 \sqrt{K} = R_0. \tag{18}$$

Denote this event by $\mathcal{I}$. In particular the reference has a nonzero initial signal component; the lemma makes no nodewise claim about the lower bound or its sign.

Proof. A sum of independent Gaussian vectors gives (17). Thus $\sqrt{d}\,\langle \bar{\boldsymbol{x}}^0, \boldsymbol{e}\rangle / \gamma_0$ is standard normal. Its density is at most $(2\pi)^{-1/2}$, so the probability that its absolute value is less than p is at most $\sqrt{2/\pi}\,p \le p$. Also $d\,\|\bar{\boldsymbol{x}}^0\|^2/\gamma_0^2 \sim \chi^2_d$ and $d\,\|\boldsymbol{X}^0\|_F^2/(K\gamma_0^2) \sim \chi^2_{Kd}$. For $Z \sim \chi^2_n$, the exponential-moment bound at parameter $3/8$ gives

$$\mathbb{P}\{Z \ge 4n\} \le \exp[-(3-\log 4)n/2] \le e^{-n/4}.$$

Apply it with $n = d$ and $n = Kd$, and take a union bound. The inequality $2\gamma_0 \le r/2$ follows from $\tau \le 1/4$ and $r_- \le r$. □

Scale convention. The initialization has covariance $\sigma_0^2 \boldsymbol{I}_d/d$ at each node and $\gamma_0^2 \boldsymbol{I}_d/d$ at the average, where $\gamma_0 = \sigma_0/\sqrt{K}$. Thus the local RMS scale is $\sigma_0 = \tau r_-$ and the initial average correlation decreases as $K^{-1/2}$. This dependence remains in T and the sample condition. Holding the average scale fixed while varying K would require changing the local scale; experimental comparisons distinguish these choices. The Gaussian RMS scale describes typical size rather than a deterministic bound, while the first mixing block controls the gradient-evaluation vectors.

C Explicit population growth and refinement

For a standard Gaussian sensing vector, fourth moments give

$$F_{\mathrm{pop}}(\boldsymbol{x}) = \tfrac{1}{4}[3\,\|\boldsymbol{x}\|^4 + 3r^4 - 2r^2\,\|\boldsymbol{x}\|^2 - 4\,\langle \boldsymbol{x}, \boldsymbol{x}_\star\rangle^2], \tag{19}$$

$$\nabla F_{\mathrm{pop}}(\boldsymbol{x}) = (3\,\|\boldsymbol{x}\|^2 - r^2)\boldsymbol{x} - 2\,\langle \boldsymbol{x}, \boldsymbol{x}_\star\rangle\,\boldsymbol{x}_\star. \tag{20}$$

The scalar dynamics in (10) follow by projecting on $\boldsymbol{e} = \boldsymbol{x}_\star/r$ and its orthogonal complement. Their validity below includes positivity of both multiplying factors.

Lemma 7 (Population recovery with an explicit horizon)**.** *Suppose* (18) *holds and* $0 < \eta r^2 \le 1/12$. *For the reference* (9) *and* T *in* (12)*, there is a fixed sign* $s_\star$ *such that*

$$\|\boldsymbol{z}^t\| \le r \quad (t \ge 0), \qquad \|\boldsymbol{z}^T - s_\star \boldsymbol{x}_\star\| \le \varepsilon r/2.$$

On the ball of radius $R = 2r_+$, ∇F_{pop} *is* $L_{\mathrm{pop}} = 39 r_+^2$*-Lipschitz.*

Proof. Orient the signal by $s_\star = \operatorname{sign}\langle \boldsymbol{z}^0, \boldsymbol{e}\rangle$ and define $a_t = \alpha_t/r > 0$, $b_t = \beta_t/r$, and $s = \eta r^2 \in (0, 1/12]$. Suppose $a_t^2 + b_t^2 \le 1$. The signal factor is $1 + 3s(1 - a_t^2 - b_t^2) \ge 1$ and the orthogonal factor is at least $1 - 2s > 0$. For $v \in [0,1]$, the function $v[1 + 3s(1-v^2)]$ is increasing since its derivative is at least $1 - 6s > 0$, and its value at 1 is 1. Bounding both factors by the signal factor proves norm invariance. Since (18) starts inside this ball, induction gives the bound for every time. It also proves preservation of the signal sign and monotonicity of a_t.

Let $q_t = b_t/a_t$ and $\log_+ v = \log\max\{1, v\}$. The two factors differ by $2s$, and hence

$$q_{t+1} = q_t\left(1 - \frac{2s}{1 + 3s(1 - a_t^2 - b_t^2)}\right) \le (1-s)q_t \le e^{-s} q_t.$$

This uses $2/(1+3s) \ge 1$. The initialization event gives

$$q_0 \le 2\sqrt{d}/p, \qquad a_0 \ge \gamma_0 p/(r\sqrt{d}).$$

Write $H = H_0$ in (12). Its definition and $s \ge s_-$ ensure that

$$sH \ge \max\{\log_+ q_0,\ \log_+(1/(2a_0)),\ \log_+(8q_0/\varepsilon),\ \log(8/\varepsilon)\}.$$

After H iterations, $q_t \le 1$. While subsequently $a_t \le 1/2$, we have $a_t^2 + b_t^2 \le 2a_t^2 \le 1/2$, so $a_{t+1} \ge (1 + 3s/2)a_t \ge e^s a_t$. The last inequality follows from $\log(1 + 3s/2) \ge s$ on $0 \le s \le 1/12$. Consequently $a_t \ge 1/2$ by time $2H + 1$, and it remains so by monotonicity. The ratio bound also implies $b_t \le \varepsilon/8$ for every $t \ge H$, since $a_t \le 1$.

Starting at $t_* = 3H + 1$, let $e_t = 1 - a_t$. The exact recurrence is

$$e_{t+1} = [1 - 3sa_t(1 + a_t)]e_t + 3sa_t b_t^2 \le (1 - 9s/4)e_t + 3s\varepsilon^2/64.$$

Here $1/2 \le a_t \le 1$, and the bracket is nonnegative because $6s \le 1/2$. Summing for H more steps yields

$$e_{4H+1} \le \tfrac{1}{2} e^{-9sH/4} + \varepsilon^2/48 \le \varepsilon/8 + \varepsilon^2/48.$$

The normalized Euclidean error is at most $e_{4H+1} + b_{4H+1} \le \varepsilon/4 + \varepsilon^2/48 < \varepsilon/2$. Thus the fixed horizon reaches the endpoint directly, without a data-dependent switch.

Finally, differentiating (20) gives

$$\nabla^2 F_{\mathrm{pop}}(\boldsymbol{x}) = 6\boldsymbol{x}\boldsymbol{x}^\top + (3\|\boldsymbol{x}\|^2 - r^2)\boldsymbol{I}_d - 2\boldsymbol{x}_\star\boldsymbol{x}_\star^\top.$$

Its operator norm is at most $9R^2 + 3r_+^2 = 39r_+^2$ on the ball. Integrating along segments proves the claimed Lipschitz bound. □

Relation to prior centralized theory. The population growth mechanism agrees with the state evolution studied by Chen et al. (2019). Their substantially sharper finite-sample result uses a different, signal-scale random start and a careful dependence analysis. Here the population argument is paired with the initialization in (5); the next uniform tensor estimate supplies the finite-data step.

D Uniform empirical fourth-moment control

The empirical gradient is cubic in the current iterate, so comparing it with its population counterpart requires fourth moments of the sensing vectors. We collect all pooled fourth-order products in the tensor $\boldsymbol{Q}$ and denote its expectation by $\boldsymbol{Q}_{\mathrm{pop}}$:

$$\boldsymbol{Q} = \frac{1}{M}\sum_{i,\ell} \boldsymbol{a}_{i,\ell}^{\otimes 4}, \qquad \boldsymbol{Q}_{\mathrm{pop}} = \mathbb{E}[\boldsymbol{a}^{\otimes 4}]. \tag{21}$$

For later comparison with matrix sensing, note that a symmetric test matrix $\boldsymbol{H}$ obeys the trace-corrected Gaussian identities

$$\mathbb{E}(\boldsymbol{a}^\top\boldsymbol{H}\boldsymbol{a})^2 = 2\|\boldsymbol{H}\|_F^2 + (\operatorname{tr}\boldsymbol{H})^2, \qquad \mathbb{E}[(\boldsymbol{a}^\top\boldsymbol{H}\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}^\top] = 2\boldsymbol{H} + \operatorname{tr}(\boldsymbol{H})\boldsymbol{I}_d. \tag{22}$$

The trace term distinguishes Gaussian phase retrieval from an isotropic matrix-sensing ensemble.

Lemma 8 (One global fourth-moment event)**.** *The tensor in* (21) *satisfies*

$$\mathbb{P}\{\|\boldsymbol{Q} - \boldsymbol{Q}_{\mathrm{pop}}\|_F > \nu\} \le \frac{d(d+2)(d+4)(d+6)}{M\nu^2}. \tag{23}$$

On the complementary event $\mathcal{Q}$, the empirical and population gradients are uniformly close over the radius-R ball:

$$\|\nabla F(\boldsymbol{x}) - \nabla F_{\mathrm{pop}}(\boldsymbol{x})\| \le \nu(\|\boldsymbol{x}\|^3 + r^2\|\boldsymbol{x}\|) \le \nu S_R. \tag{24}$$

Under (16)*, $\mathbb{P}(\mathcal{Q}^c) \le \zeta/2$.*

Proof. Treat order-four tensors as vectors with the Frobenius inner product. Independence and centering cancel the cross terms in the second moment of their sample mean:

$$\mathbb{E}\|\boldsymbol{Q} - \boldsymbol{Q}_{\mathrm{pop}}\|_F^2 = \frac{1}{M}\mathbb{E}\left\|\boldsymbol{a}^{\otimes 4} - \boldsymbol{Q}_{\mathrm{pop}}\right\|_F^2 \le \frac{1}{M}\mathbb{E}\|\boldsymbol{a}\|^8 = \frac{d(d+2)(d+4)(d+6)}{M}.$$

Markov's inequality applied to the squared norm proves (23). If $\boldsymbol{T} = \boldsymbol{Q} - \boldsymbol{Q}_{\mathrm{pop}}$, the gradient difference is the contraction of $\boldsymbol{T}$ with $\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}$ minus its contraction with $\boldsymbol{x}, \boldsymbol{x}_\star, \boldsymbol{x}_\star$. Cauchy–Schwarz for tensor contraction gives

$$\|\nabla F(\boldsymbol{x}) - \nabla F_{\mathrm{pop}}(\boldsymbol{x})\| \le \|\boldsymbol{T}\|_F\,(\|\boldsymbol{x}\|^3 + \|\boldsymbol{x}\|\,\|\boldsymbol{x}_\star\|^2).$$

This deterministic inequality holds for every $\boldsymbol{x}$, including data-dependent iterates. It gives (24) uniformly along the trajectory, so the fixed data can be reused at every step. □

Quadratic-form interpretation. For any symmetric matrix $\boldsymbol{H}$, the same event $\mathcal{Q}$ implies

$$\left| \frac{1}{M} \sum_{i,\ell} (\boldsymbol{a}_{i,\ell}^\top \boldsymbol{H} \boldsymbol{a}_{i,\ell})^2 - \left(2 \|\boldsymbol{H}\|_F^2 + (\operatorname{tr} \boldsymbol{H})^2 \right) \right| \le \nu \|\boldsymbol{H}\|_F^2 . \tag{25}$$

Indeed the difference equals $\langle \boldsymbol{Q} - \boldsymbol{Q}_{\mathrm{pop}}, \boldsymbol{H} \otimes \boldsymbol{H} \rangle$ with the matching index pairing, and $\|\boldsymbol{H} \otimes \boldsymbol{H}\|_F = \|\boldsymbol{H}\|_F^2$. Tensor Cauchy–Schwarz proves the claim simultaneously for all $\boldsymbol{H}$, hence also for low-rank matrices. This is a pooled, trace-corrected sensing bound. The proof uses this tensor event rather than an augmented matrix-sensing in-network RIP. The Frobenius estimate is stronger than the restricted control needed by many recovery arguments, which contributes to the conservative sample condition.

There is also a simple obstruction to a uniform near-linear comparison over every rank-one direction. Set $\boldsymbol{u} = \boldsymbol{a}_1 / \|\boldsymbol{a}_1\|$ and $\boldsymbol{H} = \boldsymbol{u}\boldsymbol{u}^\top$. The sample containing $\boldsymbol{a}_1$ alone gives

$$\frac{1}{M} \sum_j (\boldsymbol{a}_j^\top \boldsymbol{H} \boldsymbol{a}_j)^2 \ge \frac{\|\boldsymbol{a}_1\|^4}{M}, \qquad 2\|\boldsymbol{H}\|_F^2 + (\operatorname{tr} \boldsymbol{H})^2 = 3. \tag{26}$$

The empirical term is typically of order d^2/M. Thus a constant-factor bound that is uniform over all rank-one directions already requires order-d^2 samples. This calculation does not rule out sharper trajectory-specific analysis; it explains why an in-network matrix-sensing RIP does not directly yield a near-linear guarantee for the present proof.

E Derivative control on a bounded ball

Lemma 9 (Uniform local bounds)**.** *For any fixed finite dataset suppose* $\max_{i,\ell} \|\boldsymbol{a}_{i,\ell}\| \le B$ *and* $\max_{i,\ell} |y_{i,\ell}| \le Y$. *Let* $b_i = K m_i / M$, $b_{\max} = \max_i b_i$. *On* $\mathcal{B}(\mathbf{0}, R)$,

$$\|\boldsymbol{g}_i(\boldsymbol{x})\| \le b_i (B^4 R^3 + Y B^2 R), \tag{27}$$

$$\|\nabla \boldsymbol{g}_i(\boldsymbol{x})\|_2 \le b_i (3B^4 R^2 + Y B^2), \tag{28}$$

$$\|\nabla \boldsymbol{g}_i(\boldsymbol{x}) - \nabla \boldsymbol{g}_i(\boldsymbol{w})\|_2 \le 6 b_i R B^4 \|\boldsymbol{x} - \boldsymbol{w}\| . \tag{29}$$

The same bounds for ∇F *hold with* b_i *replaced by* 1.

Proof. The gradient is (3), and direct differentiation yields

$$\nabla \boldsymbol{g}_i(\boldsymbol{x}) = \frac{K}{M} \sum_{\ell=1}^{m_i} [3(\boldsymbol{a}_{i,\ell}^\top \boldsymbol{x})^2 - y_{i,\ell}] \boldsymbol{a}_{i,\ell} \boldsymbol{a}_{i,\ell}^\top .$$

Bound $|\boldsymbol{a}^\top \boldsymbol{x}|$ by BR and $\left\|\boldsymbol{a}\boldsymbol{a}^\top\right\|_2$ by B^2 to obtain the first two claims. For two points in the ball,

$$|(\boldsymbol{a}^\top \boldsymbol{x})^2 - (\boldsymbol{a}^\top \boldsymbol{w})^2| = |\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{w})| |\boldsymbol{a}^\top (\boldsymbol{x} + \boldsymbol{w})| \le 2B^2 R \|\boldsymbol{x} - \boldsymbol{w}\| .$$

Multiply by $3B^2$ and sum to prove (29). For F, replace K/M by $1/M$ and sum over all M samples. Integrating the Hessian along a segment in the convex ball gives the corresponding gradient Lipschitz bound; all derivative bounds here are local to this ball. □

Lemma 10 (A single Gaussian norm event)**.** *For* B *in* (14)*, the event* $\mathcal{A} = \{\max_{i,\ell} \|\boldsymbol{a}_{i,\ell}\| \le B\}$ *has probability at least* $1 - \zeta/2$. *On this event, noiseless observations obey* $y_{i,\ell} \le B^2 r_+^2$, *and the constants* L, G *in* (14) *bound every local Hessian and gradient on* $\mathcal{B}(\mathbf{0}, R)$. *Also* $H(\boldsymbol{x}) \le G$ *on this ball.*

Proof. For a standard Gaussian vector, Gaussian concentration for the Euclidean norm, with $\mathbb{E}\|\boldsymbol{a}\| \le \sqrt{d}$, gives $\mathbb{P}(\|\boldsymbol{a}\| > \sqrt{d} + \sqrt{2u}) \le e^{-u}$. Take $u = \log(2M/\zeta)$ and union bound over the M samples. Then $y \le \|\boldsymbol{a}\|^2 r^2 \le B^2 r_+^2$. Apply Lemma 9 with $b_{\max} \le K$. Finally $\boldsymbol{I}_K - \boldsymbol{J}$ is an orthogonal projection, so

$$H(\boldsymbol{x}) \le K^{-1/2} \left\|\boldsymbol{G}(\mathbf{1}\boldsymbol{x}^\top)\right\|_F \le \left(K^{-1} \sum_i G^2 \right)^{1/2} = G.$$

This event is uniform over the entire ball, all nodes, and all iterations, so it covers every data-dependent path in the proof. □

The coarse bound $b_{\max} \le K$ can be replaced by the known $b_{\max}$ in all deterministic formulas. It equals 1 for equal batches and may approach K for strongly imbalanced batches. This improves the sufficient budget using only the observed batch profile.

F NETWORK COMPARISON AND PROOF OF RECOVERY

For this section set $\boldsymbol{P} = \boldsymbol{W}^q$, $\boldsymbol{N} = \boldsymbol{I}_K - \boldsymbol{J}$, and $\delta = \rho^q$. Symmetry and double stochasticity give

$$\boldsymbol{P}\boldsymbol{J} = \boldsymbol{J}\boldsymbol{P} = \boldsymbol{J}, \qquad \|\boldsymbol{P} - \boldsymbol{J}\|_2 = \delta, \qquad \|\boldsymbol{P}\|_2 = 1. \tag{30}$$

These identities allow negative nontrivial eigenvalues of $\boldsymbol{W}$. Its nonnegative entries also imply that mixing preserves every convex ball containing the current node states.

F.1 THE AVERAGE AND DISAGREEMENT RECURSIONS

The averages of $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ coincide. Thus

$$\bar{\boldsymbol{x}}^{t+1} = \bar{\boldsymbol{x}}^t - \frac{\eta}{K}\sum_i \boldsymbol{g}_i(\boldsymbol{y}_i^t), \qquad \Delta_t^Y \le \delta\Delta_t.$$

Assume for this step that all $\boldsymbol{y}_i^t$ and the reference $\boldsymbol{z}^t$ lie in the radius-R ball. So does their average. Insert $\nabla F(\bar{\boldsymbol{x}}^t) = K^{-1}\sum_i \boldsymbol{g}_i(\bar{\boldsymbol{x}}^t)$ and the population gradient to obtain, on $\mathcal{Q}$,

$$\begin{aligned} a_{t+1} &\le \left\|\bar{\boldsymbol{x}}^t - \boldsymbol{z}^t - \eta[\nabla F_{\mathrm{pop}}(\bar{\boldsymbol{x}}^t) - \nabla F_{\mathrm{pop}}(\boldsymbol{z}^t)]\right\| \\ &\quad + \frac{\eta}{K}\sum_i \left\|\boldsymbol{g}_i(\boldsymbol{y}_i^t) - \boldsymbol{g}_i(\bar{\boldsymbol{x}}^t)\right\| + \eta\left\|\nabla F(\bar{\boldsymbol{x}}^t) - \nabla F_{\mathrm{pop}}(\bar{\boldsymbol{x}}^t)\right\| \\ &\le A_{\mathrm{pop}}a_t + \eta L\Delta_t^Y + \eta\nu S_R. \end{aligned} \tag{31}$$

The second term uses Cauchy–Schwarz and local Lipschitz continuity on the convex ball; global smoothness is unnecessary.

For disagreement subtract the consensus mean inside the second mixing block:

$$\boldsymbol{N}\boldsymbol{X}^{t+1} = (\boldsymbol{P} - \boldsymbol{J})[\boldsymbol{N}\boldsymbol{Y}^t - \eta\boldsymbol{N}\boldsymbol{G}(\boldsymbol{Y}^t)].$$

Insert $\boldsymbol{G}(\boldsymbol{1}(\bar{\boldsymbol{x}}^t)^\top)$ to obtain

$$\begin{aligned} \sqrt{K}\Delta_{t+1} &\le \delta\big[\sqrt{K}\Delta_t^Y + \eta\left\|\boldsymbol{G}(\boldsymbol{Y}^t) - \boldsymbol{G}(\boldsymbol{1}(\bar{\boldsymbol{x}}^t)^\top)\right\|_F + \eta\left\|\boldsymbol{N}\boldsymbol{G}(\boldsymbol{1}(\bar{\boldsymbol{x}}^t)^\top)\right\|_F\big] \\ &\le \sqrt{K}\delta[(1+\eta L)\Delta_t^Y + \eta H(\bar{\boldsymbol{x}}^t)]. \end{aligned} \tag{32}$$

This proves (31)–(32). Unequal batches are handled through $h_i = Km_if_i/M$ throughout.

F.2 SIMULTANEOUS CONTAINMENT AND ERROR BOUNDS

Work on $\mathcal{I}\cap\mathcal{Q}\cap\mathcal{A}$, where $\mathcal{A}$ is the sensing-norm event of Lemma 10. Lemma 7 gives $\|\boldsymbol{z}^t\| \le r \le R/2$. From (15),

$$\delta \le \min\{1/2, D/R_0, D/C_D\}. \tag{33}$$

For $\boldsymbol{W} = \boldsymbol{J}$, this is automatic since $\delta = 0$.

At $t = 0$, the mean agrees with the reference exactly, so $a_0 = 0$. The first regular mixing block gives

$$\Delta_0^Y \le \delta\frac{\left\|\boldsymbol{N}\boldsymbol{X}^0\right\|_F}{\sqrt{K}} \le \delta R_0 \le D.$$

Consequently each $\boldsymbol{y}_i^0$ is at distance at most $\sqrt{K}D \le h/2$ from $\boldsymbol{z}^0$ and lies in the radius-R ball. The unmixed starts may therefore remain outside this ball.

Inductively suppose all required gradient-evaluation states through time t lie in the ball, and $\Delta_t^Y \le D$. The recursions just proved are then justified and imply

$$\Delta_{t+1} \le \delta[(1+\eta L)D + \eta G] = \delta C_D \le D.$$

Starting from $a_0 = 0$ and summing the average recursion through time $t+1$ gives

$$a_{t+1} \le \frac{LD + \nu S_R}{L_{\mathrm{pop}}}(A_{\mathrm{pop}}^{t+1} - 1) \le \frac{h}{2A_{\mathrm{pop}}^{T}}(A_{\mathrm{pop}}^{t+1} - 1) \le h/2. \tag{34}$$

Both inequalities use the chosen D and ν and hold on the proved event $\mathcal{Q}$. Therefore

$$\left\|\boldsymbol{x}_i^{t+1} - \boldsymbol{z}^{t+1}\right\| \le a_{t+1} + \sqrt{K}\Delta_{t+1} \le h, \qquad \left\|\boldsymbol{x}_i^{t+1}\right\| \le R/2 + h \le R.$$

Since $\boldsymbol{P}$ is a nonnegative stochastic matrix, the next first mixing block preserves containment, and $\Delta_{t+1}^{Y} \le \delta\Delta_{t+1} \le D$. This closes the induction through T. In particular it proves the region assumptions used at every prior step. Intermediate post-gradient vectors are communicated only; the derivative bounds apply to the premixed evaluation states.

At the endpoint, one fixed sign from the population lemma gives

$$\left\|\boldsymbol{x}_i^{T} - s_\star \boldsymbol{x}_\star\right\| \le \left\|\boldsymbol{x}_i^{T} - \boldsymbol{z}^{T}\right\| + \left\|\boldsymbol{z}^{T} - s_\star \boldsymbol{x}_\star\right\| \le h + \varepsilon r/2 \le \varepsilon r.$$

This bounds the common-sign RMS error as well as the worst node error; the actual disagreement is at most D. Execution uses only local data and neighbor communication; the unknown truth enters the evaluator, not the algorithm. The three event bounds give total failure at most $p + 2e^{-d/4} + \zeta/2 + \zeta/2$, proving Theorem 1. The dataset is fixed and reused for all iterations. □

F.3 Budget simplification and statistical limitations

The tensor bound gives $M \gtrsim d^4 A_{\mathrm{pop}}^{2T}/(\zeta\varepsilon^2)$ because $d(d+2)(d+4)(d+6) = O(d^4)$ and $\nu^{-2} = O((r_+/h)^2 A_{\mathrm{pop}}^{2T})$. For fixed c_0, κ_0, T is logarithmic in $Kd/(\tau p\varepsilon)$, so this is polynomial but far from the near-linear guarantee of Chen et al. (2019).

For communication, $\eta L = 13 c_0 K B^4$, $G/L = 10 r_+/13$, and $L_{\mathrm{pop}} = 39 r_+^2$. Hence

$$\log \max\{2, R_0/D, C_D/D\} = O(T + \log\{K(d + \log(M/\zeta))\} + \log(1/\varepsilon)).$$

The final logarithm is absorbed into T, giving the mixing budget stated after Theorem 1. The ceiling adds at most one round, and $-\log\bar{\rho} \ge 1 - \bar{\rho}$ gives a spectral-gap form. Each iteration uses two mixing blocks, hence $2qT$ rounds, $4dqT$ words per undirected edge, and $4|\mathcal{E}|dqT$ words over the network. Optional count flooding uses scalar metadata only.

What would improve the sample bound. A norm bound for the full fourth-order tensor pays for directions the algorithm may never visit. The further amplification factor A_{pop}^{T} also treats the population map as potentially expanding at every step, including refinement. Replacing these with trajectory-specific concentration and stability is the missing ingredient for a near-linear bound. A network leave-one-out construction would have to delete a sample at its owner and propagate this change through both mixing blocks at every subsequent iteration. Here the initialization and parameter schedule are data-independent, which simplifies that construction; developing the corresponding uniform closeness proof is left for future work.

G Local interpolation and exact convergence

For Proposition 4, define the weight $\omega = \max\{1, \sqrt{K}, 4L/\mu\}$. The quantity $V_t = \left\|\bar{\boldsymbol{x}}^t - \boldsymbol{x}^\dagger\right\| + \omega\Delta_t$ combines optimization error of the network average with disagreement. A sufficient mixing requirement is

$$\rho^q \le \min\left\{\frac{\mu}{2\omega L}, \frac{1}{4(1+\eta_{\mathrm{loc}}L)}\right\}. \tag{35}$$

The first term prevents disagreement from overwhelming the curvature of the pooled loss; the second contracts the disagreement introduced by each local step.

G.1 Proof of Proposition 4

Let $a_t = \left\|\bar{\boldsymbol{x}}^t - \boldsymbol{x}^\dagger\right\|$ and $\delta = \rho^q$. Both averages and the premixed vectors $\boldsymbol{y}_i^t$ remain in the ball whenever the current node states do, because the weights are nonnegative. Integrating $\nabla^2 F$ along the

segment from $\boldsymbol{x}^\dagger$ to $\bar{\boldsymbol{x}}^t$ gives a symmetric matrix with eigenvalues in $[\mu, L]$. Since $\eta_{\mathrm{loc}} \le 1/L$ and $\nabla F(\boldsymbol{x}^\dagger) = \mathbf{0}$, centralized contraction toward this fixed point holds:

$$\left\|\bar{\boldsymbol{x}}^t - \eta_{\mathrm{loc}} \nabla F(\bar{\boldsymbol{x}}^t) - \boldsymbol{x}^\dagger\right\| \le (1 - \eta_{\mathrm{loc}}\mu) a_t.$$

Interpolation and Lipschitz continuity also give

$$H(\bar{\boldsymbol{x}}^t) \le K^{-1/2} \left\|\boldsymbol{G}(\mathbf{1}(\bar{\boldsymbol{x}}^t)^\top)\right\|_F \le L a_t.$$

The mean-gradient discrepancy is at most $\eta_{\mathrm{loc}} L \Delta_t^Y \le \eta_{\mathrm{loc}} L \delta \Delta_t$. The second mixing block gives disagreement at most $\delta[(1 + \eta_{\mathrm{loc}} L)\delta\Delta_t + \eta_{\mathrm{loc}} L a_t]$. Bounding $\delta \le 1$ yields the conservative componentwise comparison

$$\begin{pmatrix} a_{t+1} \\ \Delta_{t+1} \end{pmatrix} \le \begin{pmatrix} 1 - \eta_{\mathrm{loc}}\mu & \eta_{\mathrm{loc}} L \\ \delta \eta_{\mathrm{loc}} L & \delta(1 + \eta_{\mathrm{loc}} L) \end{pmatrix} \begin{pmatrix} a_t \\ \Delta_t \end{pmatrix}. \tag{36}$$

For $V_t = a_t + \omega \Delta_t$, its a_t coefficient is at most $1 - \eta_{\mathrm{loc}}\mu + \omega\delta\eta_{\mathrm{loc}} L \le 1 - \eta_{\mathrm{loc}}\mu/2$. Its Δ_t coefficient is bounded by

$$\eta_{\mathrm{loc}} L + \omega\delta(1 + \eta_{\mathrm{loc}} L) \le \omega(\eta_{\mathrm{loc}}\mu/4 + 1/4) \le \omega/2 \le \omega(1 - \eta_{\mathrm{loc}}\mu/2),$$

because $\eta_{\mathrm{loc}}\mu \le 1$. Thus $V_{t+1} \le (1 - \eta_{\mathrm{loc}}\mu/2) V_t$ whenever the current nodes lie in the ball. Finally

$$\left\|\boldsymbol{x}_i^t - \boldsymbol{x}^\dagger\right\| \le a_t + \sqrt{K}\Delta_t \le V_t.$$

The condition $V_0 \le b$ establishes containment initially, and $V_{t+1} \le V_t \le b$ closes the induction. This proves (11) for all time. The contraction is toward one fixed common solution, which is the quantity controlled by this proposition. □

G.2 Verifiable existence of a local regular region

For a noiseless dataset let $B = \max_{i,\ell} \|\boldsymbol{a}_{i,\ell}\|$ and

$$\boldsymbol{H}_\star = \nabla^2 F(\boldsymbol{x}_\star) = \frac{2}{M} \sum_{i,\ell} y_{i,\ell} \boldsymbol{a}_{i,\ell} \boldsymbol{a}_{i,\ell}^\top, \qquad \lambda_\star = \lambda_{\min}(\boldsymbol{H}_\star). \tag{37}$$

If $\lambda_\star > 0$, take

$$b = \min\{r, \lambda_\star/(24 r B^4)\}, \quad \mu = \lambda_\star/2, \quad L = 13 b_{\max} B^4 r^2.$$

Every point in $\mathcal{B}(\boldsymbol{x}_\star, b)$ has norm at most $2r$. Lemma 9, applied to F with $R = 2r$, implies

$$\left\|\nabla^2 F(\boldsymbol{x}) - \boldsymbol{H}_\star\right\|_2 \le 12 r B^4 \|\boldsymbol{x} - \boldsymbol{x}_\star\| \le \lambda_\star/2.$$

Thus $\nabla^2 F(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_d$. The local Hessian norms are at most $13 b_{\max} B^4 r^2$, which also bounds the global Hessian because $b_{\max} \ge 1$. Interpolation holds automatically. Reflection gives the same properties around $-\boldsymbol{x}_\star$.

For a fixed nonzero $\boldsymbol{x}_\star$ and Gaussian data, $y_{i,\ell} > 0$ almost surely. If $M \ge d$, the sensing vectors span $\mathbb{R}^d$ almost surely. Therefore, for any nonzero $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \boldsymbol{H}_\star \boldsymbol{v} = \frac{2}{M} \sum_{i,\ell} y_{i,\ell} (\boldsymbol{a}_{i,\ell}^\top \boldsymbol{v})^2 > 0,$$

so $\lambda_\star > 0$ almost surely. This proves existence of a local region, while its radius and curvature depend on the realized data and truth. Turning the local result into a global two-stage algorithm would require a computable switch and certified curvature bounds; the finite-accuracy theorem instead uses a fixed horizon.

Noise separates three errors. For noisy data, let $\boldsymbol{x}_{\mathrm{emp}}$ denote a well-defined centralized empirical solution in a chosen basin, if such a solution exists. Then one may distinguish the statistical error $\mathrm{dist}(\boldsymbol{x}_{\mathrm{emp}}, \boldsymbol{x}_\star)$, the optimization error of the average relative to that solution, and disagreement. If $b_g = K^{-1/2} \left\|\boldsymbol{G}(\mathbf{1}\boldsymbol{x}_{\mathrm{emp}}^\top)\right\|_F > 0$, the second row of (36) acquires an additive term at most $\delta\eta_{\mathrm{loc}} b_g$. This identifies a possible bias mechanism; a noisy theorem would additionally require basin curvature, stability, and statistical concentration.

H PRIOR-WORK COMPARISON AND ALGORITHM DISTINCTIONS

Table 1 summarizes the scope of selected primary sources. The table records which guarantees apply to the present finite-sample planted-signal model; a different scope does not imply that a method fails to recover signals. The comparison is selective rather than an exhaustive novelty certificate. Decentralized phase retrieval itself predates this work.

Table 1: Scope relevant to this manuscript. Additional assumptions in each source remain necessary.

Work	Architecture and initialization	Guarantee and distinction
Chen et al. (2019)	Centralized; random Gaussian vector or empirical scale times random direction	Gaussian fixed data; near-linear samples up to logs; finite-time recovery. Used as a benchmark for the present small starts.
Zhao et al. (2018)	Peer-to-peer primal–dual DWF; feasible starting point	Approximate KKT convergence at a sublinear rate; its guarantee concerns optimization rather than random-start planted recovery.
Wang et al. (2020)	Partitioned expectation-consistent inference	Noisy phase retrieval and state evolution; its architecture differs from arbitrary connected gossip graphs.
Mafakheri et al. (2024)	Peer-to-peer asynchronous ADMM	Algorithmic relation to randomized block Douglas–Rachford splitting; phase retrieval application. Different update and target guarantee.
Yang (2025)	Multi-view sparse sensing; global amplitude stage at a server	Outlier and view-structure recovery conditions; a server-based shared-vector model.
Maros & Scutari (2023)	Peer-to-peer matrix sensing; independent small Gaussian factor initialization	In-network RIP, alignment and refinement; sensing ensemble differs from rank-one intensities.
This draft	Mix–gradient–mix; independent Gaussian vectors; native initialization scale	Explicit but conservative polynomial sample bound; finite-accuracy common-sign recovery; conditional local exactness.

Choice of method. One gossip step per gradient is simple but may leave a large early perturbation. Mixing before and after the gradient controls both the evaluation points and the injected disagreement, at the cost of an additional communication block. Gradient tracking can correct persistent gradient heterogeneity by carrying an auxiliary vector and a coupled error. We use the mix–gradient–mix option with q rounds in each block, controlling gradient-evaluation disagreement from the first iteration onward. The proof therefore focuses on this update. Overparameterization and preconditioning do not enter the analysis. In rank-one recovery there is no nontrivial ratio between multiple nonzero signal singular values, although the empirical loss Hessian can have a substantial condition number.

H.1 WHY ELIMINATING THE DWF DUAL DOES NOT PRODUCE GOSSIP-PR

The original draft's distinction between the two algorithms is retained here with consistent notation. Let $\boldsymbol{B}$ be a signed edge-node incidence matrix, $\boldsymbol{D}$ the degree matrix, and $\boldsymbol{A}_{\mathcal{G}}$ the adjacency matrix. Assume $K \geq 2$ and a connected unweighted graph, and define

$$\boldsymbol{L} = \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{D} - \boldsymbol{A}_{\mathcal{G}}, \qquad \boldsymbol{P} = \tfrac{1}{2}\boldsymbol{D}^{-1}(\boldsymbol{D} + \boldsymbol{A}_{\mathcal{G}}), \qquad \boldsymbol{S} = (2\beta \boldsymbol{D})^{-1}, \quad \theta = 1 - \beta\gamma_{\mathrm{d}} \in (0, 1).$$

After restricting the published DWF update to real variables and matching its objective scaling, it has the form

$$\boldsymbol{X}^{t+1} = \mathrm{Proj}_{\mathcal{C}^K}[\boldsymbol{P}\boldsymbol{X}^t - \boldsymbol{S}\boldsymbol{G}(\boldsymbol{X}^t) - \theta \boldsymbol{S}\boldsymbol{B}^\top \boldsymbol{\Lambda}^t], \tag{38}$$

$$\boldsymbol{\Lambda}^{t+1} = \theta \boldsymbol{\Lambda}^t + \beta \boldsymbol{B}\boldsymbol{X}^{t+1}. \tag{39}$$

Here $\gamma_{\mathrm{d}} > 0$ is dual damping, unrelated to the initialization scale γ_0. The real linear gradient term in the source's equation (12) has coefficient $2/M$, while our $\boldsymbol{g}_i$ has coefficient K/M. Write $c_f = 2/K$. Dividing its primal optimality equation by c_f gives $\beta = \beta_{\mathrm{pub}}/c_f$, $\gamma_{\mathrm{d}} = c_f\gamma_{\mathrm{pub}}$, and $\boldsymbol{\Lambda} = \boldsymbol{\omega}_{\mathrm{pub}}/c_f$; the product defining θ is unchanged. The displayed updates therefore match after rescaling, while the source's parameter ranges remain separate.

Proposition 11 (Primal-only form of DWF)**.** *If projections in* (38) *are inactive, then for* $t \geq 1$*,*

$$\boldsymbol{X}^{t+1} = (1+\theta)\boldsymbol{P}\boldsymbol{X}^t - \theta\boldsymbol{P}\boldsymbol{X}^{t-1} - \boldsymbol{S}[\boldsymbol{G}(\boldsymbol{X}^t) - \theta\boldsymbol{G}(\boldsymbol{X}^{t-1})]. \tag{40}$$

Conversely, this recurrence reproduces the same primal path when its first two iterates are those generated by the chosen $\boldsymbol{X}^0, \boldsymbol{\Lambda}^0$ *in* (38)*.*

Proof. At time $t-1$, rearrange the inactive primal equation to obtain

$$\theta\boldsymbol{S}\boldsymbol{B}^\top\boldsymbol{\Lambda}^{t-1} = \boldsymbol{P}\boldsymbol{X}^{t-1} - \boldsymbol{S}\boldsymbol{G}(\boldsymbol{X}^{t-1}) - \boldsymbol{X}^t.$$

Multiply the dual update by $\theta\boldsymbol{S}\boldsymbol{B}^\top$ and substitute:

$$\theta\boldsymbol{S}\boldsymbol{B}^\top\boldsymbol{\Lambda}^t = \theta\boldsymbol{P}\boldsymbol{X}^{t-1} - \theta\boldsymbol{S}\boldsymbol{G}(\boldsymbol{X}^{t-1}) - \theta\boldsymbol{X}^t + \theta\beta\boldsymbol{S}\boldsymbol{L}\boldsymbol{X}^t.$$

Insert into the primal equation and use $\beta\boldsymbol{S}\boldsymbol{L} = \boldsymbol{I}_K - \boldsymbol{P}$. This proves (40). Starting from the matched first step, reconstruction of the dual sequence through (39) and induction prove the converse. □

If projections are active, their normal-cone terms $\boldsymbol{R}^{t+1}$ enter the primal equation as $-\boldsymbol{S}\boldsymbol{R}^{t+1}$. Elimination then adds $-\boldsymbol{S}(\boldsymbol{R}^{t+1} - \theta\boldsymbol{R}^t)$ to (40), so the primal-only recurrence applies only when these terms vanish. On a regular graph the formal limit $\theta = 1$ has an EXTRA-type two-step form (Shi et al., 2015); this limit lies outside the positive-damping setting above. Equivalence and convergence therefore require the matched initialization and the stated projection regime.

In contrast, GOSSIP-PR is the one-step map $\boldsymbol{X}^+ = \boldsymbol{W}^{2q}\boldsymbol{X} - \eta\boldsymbol{W}^q\boldsymbol{G}(\boldsymbol{W}^q\boldsymbol{X})$. On a regular graph with $\boldsymbol{W}^q = \boldsymbol{P}$, $\boldsymbol{S} = \eta\boldsymbol{I}_K$, a consensus initial state $\boldsymbol{X}^0 = \mathbf{1}(\boldsymbol{x}^0)^\top$, and $\boldsymbol{\Lambda}^0 = \mathbf{0}$, the first-step difference is

$$\boldsymbol{X}^1_{\mathrm{DWF}} - \boldsymbol{X}^1_{\text{GOSSIP-PR}} = -\eta(\boldsymbol{I}_K - \boldsymbol{P})\boldsymbol{G}(\boldsymbol{X}^0).$$

The difference is generally nonzero. For example, take two nodes with $d = m_1 = m_2 = 1$, $x_\star = 1$, sensing values $a_{1,1} = 1$, $a_{2,1} = 2$, and a common initial value $1/2$. Then $\boldsymbol{G}(\boldsymbol{X}^0) = (-0.375, -6)^\top$. With $\eta = 0.01$, $\boldsymbol{P} = \boldsymbol{J}$, and inactive projection, the DWF iterate is $(0.50375, 0.56)^\top$, whereas GOSSIP-PR gives $(0.531875, 0.531875)^\top$. This exact arithmetic example isolates the algebraic identity.

At a fixed point with $0 < \theta < 1$, the interior DWF equations imply $\boldsymbol{B}\boldsymbol{X} = \gamma_{\mathrm{d}}\boldsymbol{\Lambda}$ and

$$\boldsymbol{G}(\boldsymbol{X}) + \gamma_{\mathrm{d}}^{-1}\boldsymbol{L}\boldsymbol{X} = \mathbf{0}.$$

Thus a finite-penalty stationary point can fail to be consensual. Both methods nevertheless share the interpolating fixed points $\boldsymbol{X} = \mathbf{1}(s\boldsymbol{x}_\star)^\top$, with zero dual state for DWF. Shared fixed points alone do not determine the trajectories or recovery guarantees.

I EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

This section gives the data generation, comparators, and evaluation rules for the reported synthetic experiments. The scripts and manifests under `experiments/` use the same configurations and seeds.

I.1 DATA GENERATION AND EVALUATION

For each trial, we drew $\boldsymbol{v}_\star \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, set $\boldsymbol{x}_\star = \boldsymbol{v}_\star / \|\boldsymbol{v}_\star\|$, and drew M standard Gaussian sensing vectors. The observations $y_j = (\boldsymbol{a}_j^\top \boldsymbol{x}_\star)^2$ and sensing vectors were fixed throughout a run. Paired methods used the same truth, data, partition, and initialization draws, but did not access $\boldsymbol{x}_\star$ during optimization.

All runs used $r = 1$, the public calibration interval $[r_-, r_+] = [1/2, 2]$, and independent Gaussian starts with $\tau = 0.1$ unless an initialization ablation states otherwise. Twenty pilot seeds beginning

at 20260901 selected step size 0.02 for GOSSIP-PR, 0.01 for gradient tracking, and $(\beta, \gamma_{\mathrm{d}}) = (20, 0.025)$ for DWF. The pilot streams were disjoint from every evaluation stream.

Independent `SeedSequence` streams generated the truth, data, topology, partition, and initialization. Pilot seeds began at 20260901, and disjoint evaluation streams were derived from 20270921. Primary recovery and cooperation cells used 100 evaluation seeds; trajectory, topology, initialization, and ablation cells used 50. The dense communication–step grid used 20 additional seeds. All computations used float64 arithmetic.

Balanced partitions assigned either $\lfloor M/K \rfloor$ or $\lceil M/K \rceil$ samples to each node. The imbalance study used geometric batch profiles with target ratios $1, 4, 16, 64$, rounded to preserve M and $m_i \geq 1$, followed by a random vertex assignment. Its comparison used the pooled objective $\sum_i (m_i/M) f_i$ and the explicitly labeled diagnostic $K^{-1} \sum_i f_i$.

The main-text success thresholds and 1200-pass budget are specified in Section 5; the initialization ablation used 1600 passes. Runs with nonfinite arithmetic or an iterate norm above $10^6 r_+$ were classified as divergent, and trials that missed the target within the budget were censored. No failed, divergent, or censored trial was removed. Success fractions include Wilson 95% intervals, while trajectory plots show medians and interquartile ranges.

I.2 COMPARATORS

Centralized random-start gradient descent minimized F from the average of the nodewise starts and used the same step as GOSSIP-PR. This empirical path is distinct from the population reference $\boldsymbol{z}^t$. Centralized spectral-start gradient descent formed $\boldsymbol{Y} = M^{-1} \sum_j y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top$, scaled its leading eigenvector by $\widehat{r} = (M^{-1} \sum_j y_j)^{1/2}$, and refined it with the same loss. Independent local baselines applied random-start or spectral-start descent using only one node's measurements; sign alignment was used only for evaluation.

DWF used the real primal–dual restriction of Zhao et al. (2018), including projection and dual damping, with the normalization in Appendix H.1. The disjoint pilot selected $(\beta, \gamma_{\mathrm{d}}) = (20, 0.025)$. The gradient-tracking comparator used

$$\boldsymbol{X}^{t+1} = \boldsymbol{W}(\boldsymbol{X}^t - \eta \boldsymbol{S}^t), \qquad \boldsymbol{S}^{t+1} = \boldsymbol{W}\boldsymbol{S}^t + \boldsymbol{G}(\boldsymbol{X}^{t+1}) - \boldsymbol{G}(\boldsymbol{X}^t), \qquad \boldsymbol{S}^0 = \boldsymbol{G}(\boldsymbol{X}^0),$$

and transmitted both state variables. The decentralized spectral baseline used 20 power iterations with four gossip rounds per iteration before the same GOSSIP-PR refinement; its preprocessing messages were included in the communication count.

I.3 RECOVERY, COOPERATION, AND DYNAMICS

The primary recovery grid and graph families are specified in Section 5.

The cooperation study fixed (d, M) at $(128, 1024)$ or $(256, 2048)$ and varied $K \in \{2, 4, 8, 16, 32, 64\}$. Paired trials repartitioned the same pooled data. We compared fixed $q = 16$ with the data-dependent choice satisfying $\rho^q \leq 10^{-2}$ and marked the regime in which every $m_i < d$.

At $(d, K) = (128, 16)$, full trajectories were retained for the successful cell $(M/d, q) = (12, 16)$ and the unsuccessful cell $(M/d, q) = (3, 1)$. The displayed components are

$$\alpha_t = |\langle \bar{\boldsymbol{x}}^t, \boldsymbol{x}_\star \rangle|, \qquad \beta_t = \left\| \bar{\boldsymbol{x}}^t - \langle \bar{\boldsymbol{x}}^t, \boldsymbol{x}_\star \rangle \boldsymbol{x}_\star \right\|,$$

and basin entry is the first time the common-sign maximum error is at most 0.1.

I.4 TOPOLOGY AND ABLATIONS

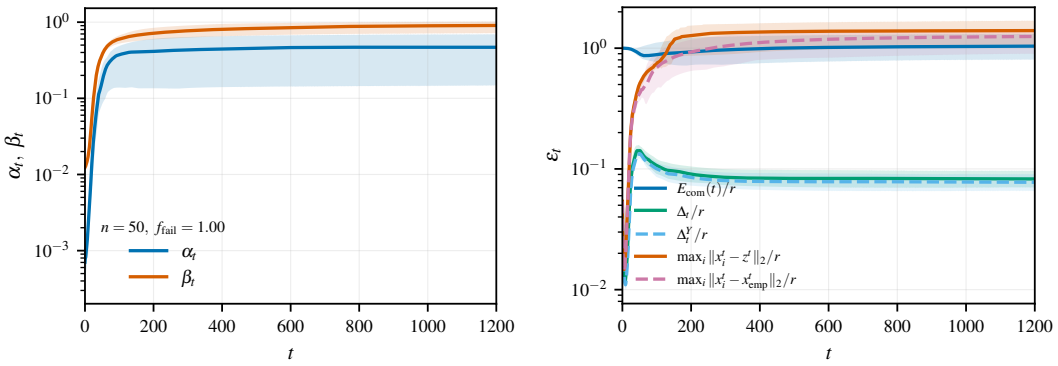
The topology study used $(d, M, K) = (128, 1024, 16)$ and $(128, 1024, 64)$ with complete graphs, cycles, square tori, connected random 4-regular graphs, and connected geometric graphs. A geometric graph used the smallest radius on a fixed grid that produced connectivity. Sparse graphs used symmetric Metropolis weights, $W_{ij} = 1/(1 + \max\{\deg(i), \deg(j)\})$ on edges and $W_{ii} = 1 - \sum_{j \neq i} W_{ij}$; complete graphs used $\boldsymbol{W} = \boldsymbol{J}$. For $q \in \{1, 2, 4, 8, 16, 32, 64\}$, the reported costs are pooled gradient passes, $2qT_{10^{-3}}$ gossip rounds, and $4|\mathcal{E}|dqT_{10^{-3}}$ transmitted words.

At $(d, M, K) = (128, 1024, 16)$, the communication–step study crossed $q \in \{1, 2, 4, 8, 16, 32\}$ with $\eta r_+^2 \in \{0.005, 0.01, 0.02, 0.04, 0.08\}$ on a cycle and a random 4-regular graph. Figure 12 uses a denser cycle grid with $q \in \{1, 2, 4, 6, 8, 12, 16, 24, 32\}$ and $\eta r_+^2 \in \{0.005, 0.01, 0.015, 0.02, 0.03, 0.04, 0.06, 0.08\}$.

The initialization study used $(d, M) = (128, 1024)$, $K \in \{4, 16, 64\}$, and $\tau \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. It compared independent and shared Gaussian starts at fixed σ_0 or fixed γ_0, a shared spherical start, centralized and decentralized spectral directions, and a half-positive/half-negative common-direction stress test. The imbalance study used $(d, M, K) = (128, 1024, 16)$ and target batch ratios $1, 4, 16, 64$.

I.5 Additional empirical results

Recovery transition and failed dynamics. The transition slices at $K = 4$ and $K = 64$ follow the same qualitative threshold as the representative $K = 16$ slice in Figure 3a and are omitted as redundant. At the deliberately difficult cell $(d, M/d, K, q) = (128, 3, 16, 1)$, none of 50 trials meets the success criterion. Figure 5a shows inconsistent signal growth and Figure 5b shows persistent network and reference errors; their median final common-sign error and disagreement are 1.04 and 8.26×10^{-2}.

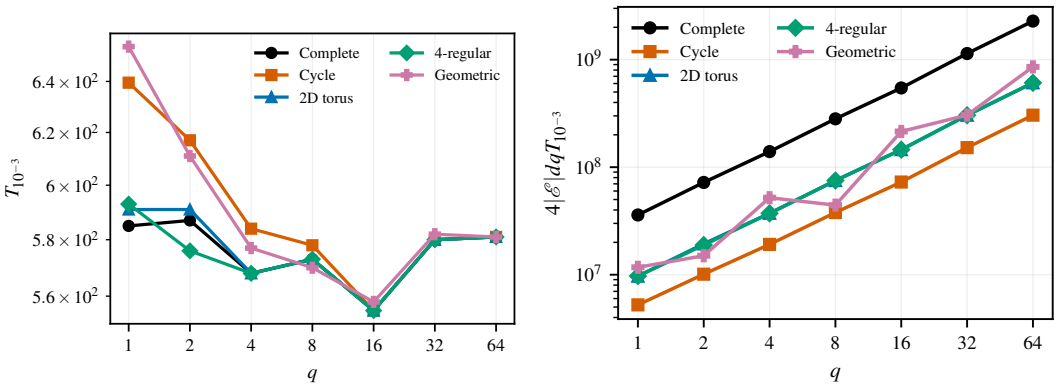


(a) Signal and orthogonal components.

(b) Network and reference errors.

Figure 5: Unsuccessful-cell dynamics over 50 seeds.

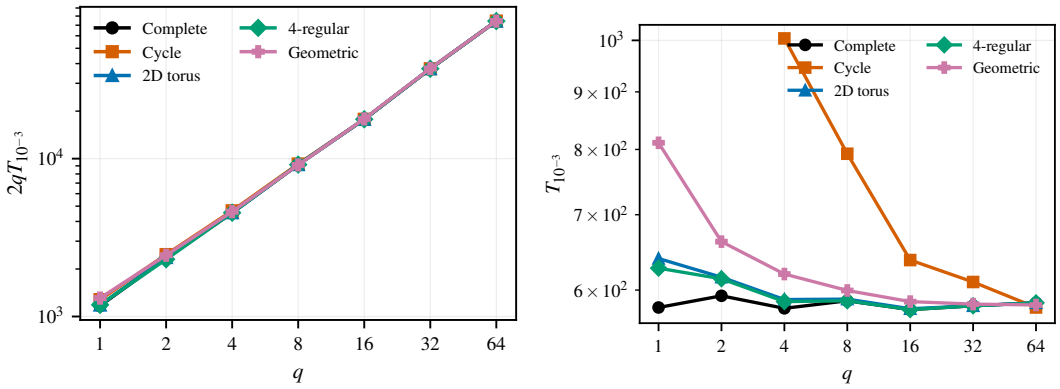
Additional topology accounting. Figures 6a, 7a, and 6b give all three target-reaching cost views at $K = 16$. Figures 7b, 8a, and 8b give the corresponding views at $K = 64$. Two of the 70 graph–q cells remain censored because no trial reaches the target within the budget.



(a) Pooled gradient passes.

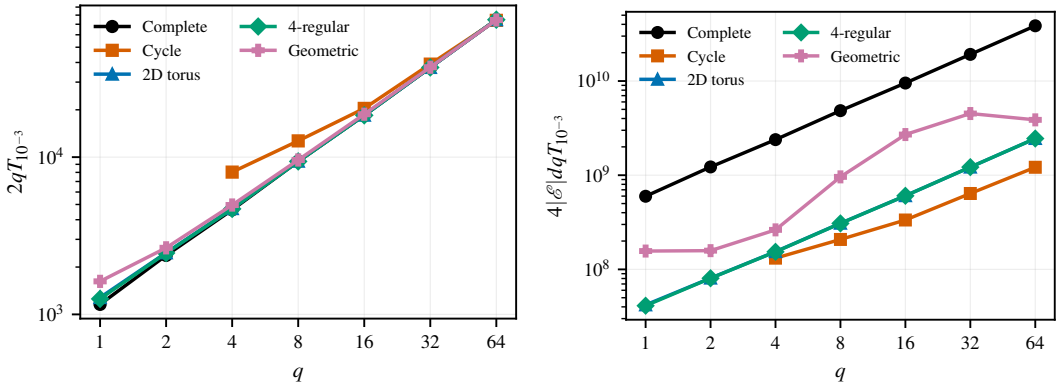
(b) Network-wide transmitted words.

Figure 6: Median target-reaching cost at $K = 16$; censored cells are omitted.



(a) $2qT_{10^{-3}}$ at $K = 16$. (b) $T_{10^{-3}}$ at $K = 64$.

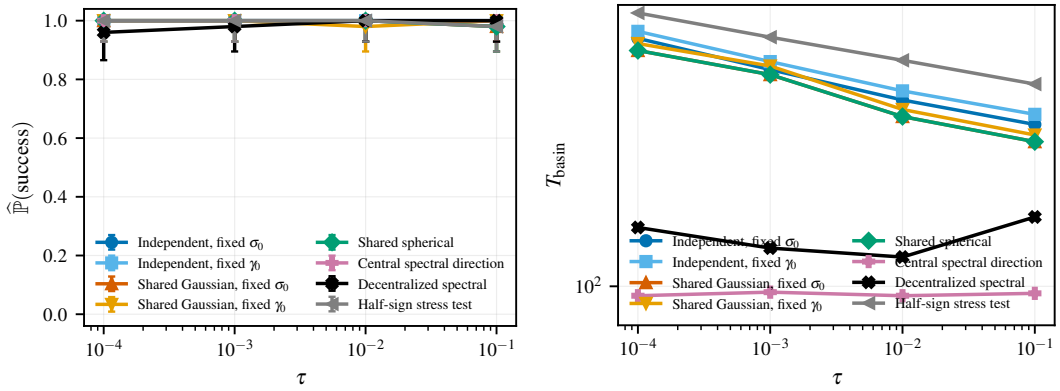
Figure 7: Additional topology costs.



(a) $2qT_{10^{-3}}$. (b) $4|\mathcal{E}|dqT_{10^{-3}}$.

Figure 8: Communication costs at $K = 64$.

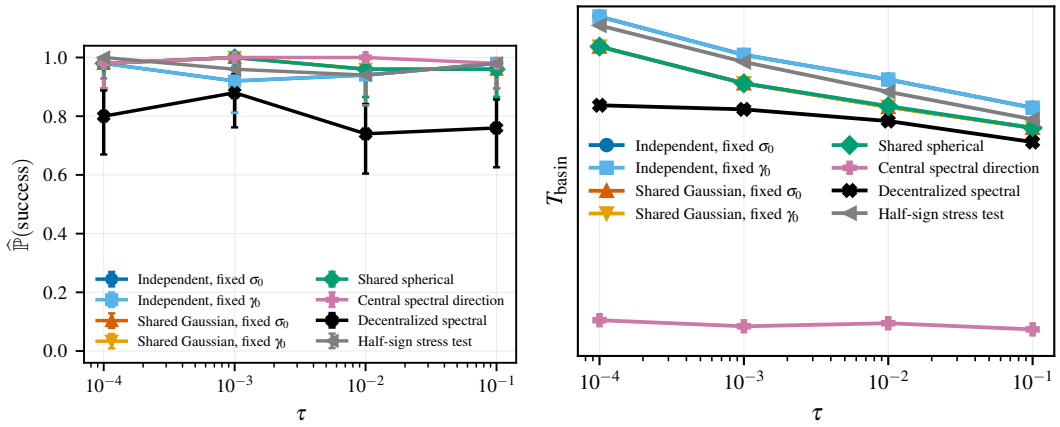
Initialization ablation. Figures 9a and 10a show final success; Figures 9b and 10b show basin-entry time. These two network sizes retain the representative and large-scale behaviors, while the qualitatively redundant $K = 4$ traces are omitted. Spectral directions enter earlier, while their preprocessing cost is charged in Figures 11a and 11b. For example, at $K = 64$ the distributed spectral setup uses 1.32×10^6 words and attains success probability 0.80 at $\tau = 10^{-4}$, compared with 0.98 for the primary independent start without setup messages.



(a) Success, $K = 16$.

(b) Basin time, $K = 16$.

Figure 9: Initialization at $K = 16$.

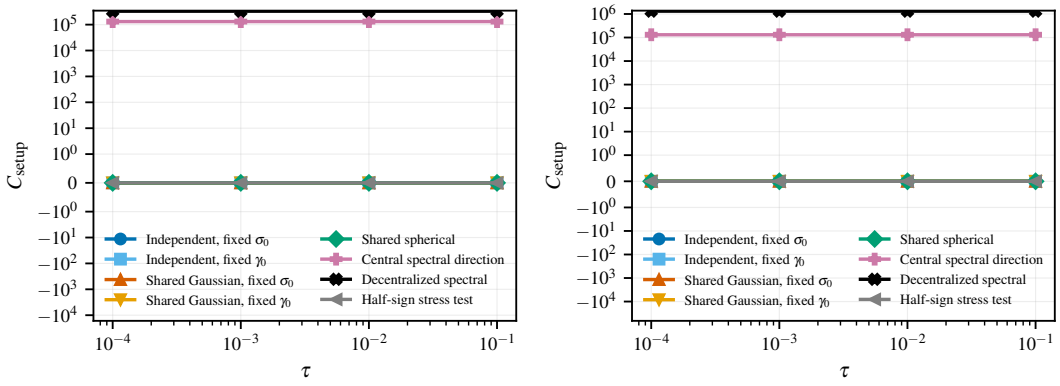


(a) Success, $K = 64$.

(b) Basin time, $K = 64$.

Figure 10: Initialization at $K = 64$.

Step size, gossip, and imbalance. Figure 12 reports the denser cycle grid. Grayscale encodes empirical success probability from black (zero) to white (one); all displayed cells have zero divergent or nonfinite runs. The random 4-regular graph gives the same qualitative pattern and is omitted. At $\eta r_+^2 \leq 0.02$, runs remain above the strict target after 1200 passes. Figure 13 compares the correct pooled objective with the labeled unweighted diagnostic. At target batch ratios 16 and 64, the correctly weighted method succeeds on 0.56 and 0.64 of trials, whereas the unweighted diagnostic has no successes. This noiseless result concerns finite-budget transients rather than asymptotic statistical bias.



(a) $K = 16$. (b) $K = 64$.

Figure 11: Initialization setup words.

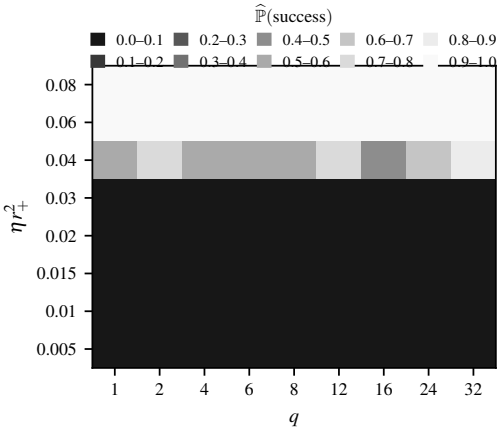


Figure 12: Success probability over a dense communication–step grid on a cycle.

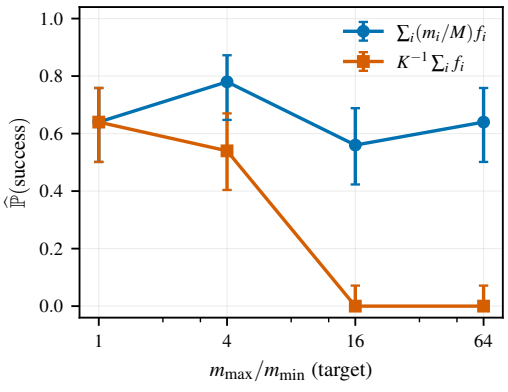


Figure 13: Sample-imbalance ablation.