ZO-COSMO: INDEX-FREE ONE-HOP MIXING FOR DECENTRALIZED ZEROTH-ORDER OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse communication in decentralized zeroth-order learning requires compatible peer-state coordinates. We characterize this one-hop condition and develop ZO-COSMO, coupling two-query estimation with average-preserving masked consensus using q values per active link. Global supports serve all-neighbor mixing; matching updates require agreement only within each pair. We derive a sharp contraction-per-scalar bound within the matching class and convergence guarantees for the core and sparse-momentum updates. At fixed matching, exact moment identities characterize how shared directions preserve gradient-heterogeneity cancellation and redistribute estimation error and disagreement. Mechanism experiments cover unequal curvatures, noise, and sparse momentum. Further tests span 64 synthetic agents and eight logical Qwen LoRA workers. At matched payload budgets, Qwen2-7B QNLI gains 3.65 accuracy points over explicit-index Rand-k; edge-local updates gain 3.42 and 2.53 points over all-neighbor mixing on eight-worker complete and ring graphs. A matched-first-step ablation gives a 3.92-point momentum benefit. Seed-aware and same-matching controls distinguish encoding, scheduling, and query correlation.

1 INTRODUCTION

In decentralized black-box learning, agents can query local function values but cannot centralize data or gradients. Let $x \in \mathbb{R}^d$ denote the shared decision vector. We study

$$\min_{x\in\mathbb{R}^d} F(x), \qquad F(x) = \frac{1}{N}\sum_{i=1}^{N} F_i(x) \tag{1}$$

over a graph, where each F_i is a smooth nonconvex black box. Zeroth-order methods replace gradients with finite differences (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017; Shamir, 2017), but high dimension makes both estimation and communication expensive.

Sparse communication helps only if the support is also cheap. A coordinate-list encoding sends $qB_{\mathrm{val}} + q\lceil \log_2 d \rceil$ bits. At billion-parameter scale, each index is about as wide as a single-precision value, nearly doubling the value-only payload.

We ask:

> ***Can a decentralized zeroth-order method communicate sparse updates without communicating sparse indices?***

Public seeds already reconstruct ZO updates at a server (Li et al., 2025; Cai et al., 2026; Ran et al., 2026) or through graph-wide dissemination (Kim & Lee, 2026). One-hop mixing also requires the corresponding peer-state values. A message carrying values on $\{2, 3\}$ leaves the value at coordinate 1 unavailable for an average on $\{1, 2\}$, even when its support is public (Figure 1). Theorem 1 gives the necessary and sufficient support condition for mixing arbitrary current peer states without stored peer histories.

ZO-COSMO meets this condition by coordinating supports with the mixing graph. A global support permits masked Metropolis mixing; a matching permits separate supports on disjoint edges and one

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.

q-value message per active node. Uniform perfect matchings on a complete graph attain the best contraction per scalar within this operator class (Theorem 3). The query and mixing steps share a mask, so their dependence also enters the optimization analysis. We show how a shared direction preserves cancellation between local gradients and how independent directions redistribute estimation error and disagreement.

1.1 CONTRIBUTIONS

- **Exact index-free mixing.** We characterize the supports needed to mix arbitrary peer states from one-hop coordinate messages, with a sharp missing-coordinate error bound. This leads to global and matching-based protocols with q values per active link.
- **Mixing and optimization mechanisms.** We derive the optimal contraction per scalar in the matching class and separate average-estimation error from disagreement injection. Pair invariance yields a heterogeneity-sensitive core bound, and exact conditional moments account for oracle noise and stored momentum; separate results establish bit-budget and sparse-momentum convergence.
- **Mechanism and model-scale evidence.** Fixed-matching experiments test correlation under unequal curvatures, noise, and momentum. Qwen LoRA shows fixed-budget gains over explicit-index and all-neighbor baselines and a matched-step momentum benefit. Seed-aware and same-matching controls separate these effects; synthetic tests include adaptive Top-k compression and 64 agents.

2 RELATED WORK

Zeroth-order and decentralized optimization. ZO methods make optimization possible when gradients are unavailable, using one- or two-point function queries (Conn et al., 2009; Flaxman et al., 2005; Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017; Shamir, 2017; Larson et al., 2019). In high dimension, much of the work has focused on better directions through sparsity, subspaces, momentum, and variance reduction (Wang et al., 2018; Golovin et al., 2019; Liu et al., 2018b;a; Cai et al., 2020; Kozak et al., 2020; Chen et al., 2019). Decentralized methods must also keep locally updated states in agreement over a graph (Yuan & Ho, 2015; Yuan et al., 2015; Sahu et al., 2018; Hajinezhad et al., 2019; Tang & Li, 2019; Zhang & Bailey, 2022; Veprikov et al., 2024). We retain two-point estimation and study which sparse messages allow that mixing step to be executed exactly.

Compressed decentralized learning. Randomized gossip relates pairwise averaging to graph spectra (Boyd et al., 2006); compressed learning uses quantization, sparsification, sign compression, and error compensation (Alistarh et al., 2017; Lin et al., 2018; Stich et al., 2018; Lan et al., 2020; Tang et al., 2018; Koloskova et al., 2019; 2020; Bernstein et al., 2018; Basu et al., 2019). Coordinate lists add a representation cost (Lin et al., 2018; Shi et al., 2020). PermK analyzes how correlated compressors change aggregate estimation error (Szlendak et al., 2022). Our query direction also selects the peer-state coordinates that are mixed. We therefore study both aggregate error and the disagreement injected by that same update. Complementary supports may reduce aggregate variance without supplying the coordinates required for exact current-state mixing.

DeComFL and MEERKAT reconstruct seed–scalar updates at a server, while FeedSign aggregates seed–sign votes (Li et al., 2025; Ran et al., 2026; Cai et al., 2026). SeedFlood disseminates updates across the graph and batches their application in a shared subspace (Kim & Lee, 2026). These methods exploit compact update representations. ZO-COSMO sends q current-state values to complete one neighbor-mixing stage; Appendix C.1 compares the payload and dissemination requirements. Recent decentralized ZO methods combine estimators with general compressors (Hua et al., 2026; Wang et al., 2025; Chen et al., 2025). We include Com-DSZO with adaptive Top-k and a periodic-refresh comparator motivated by MARINA (Gorbunov et al., 2021).

Forward-only model adaptation. Language-model adaptation can replace backpropagation with loss queries (Malladi et al., 2023; Zhang et al., 2024; Liu et al., 2025; Yu et al., 2025). Sparse and low-dimensional directions reduce the parameters touched by each query, but peers must still agree on how received values update their local states. Our Qwen experiments examine this communication step with separate worker states and data shards.

3 ALGORITHM

Algorithm 1 reconstructs a public support, takes two sparse queries, and mixes selected coordinates. Messages contain q ordered values; unsent coordinates retain their local values. The implementation derives $R_t = \mathcal{H}(S_{\text{master}} \,\|\, t)$ from synchronized seeds and round counters. Decoding is pathwise exact; stochastic rates use fresh public coins (Appendix A.4).

Let X_t and Y_t stack the node states and pre-mixing updates as rows. The network step has the compact form

$$X_{t+1} = Y_t(I - P_t) + WY_tP_t. \tag{2}$$

Theorem 1 determines which coordinates messages must cover. ZO-COSMO uses this support for both queries and mixing, coupling local progress to the communicated coordinates.

A global support is sufficient but not necessary. In Algorithm 2, a public matching assigns each edge its own support. Endpoints query and average on this support, using one q-value message per active node; disjoint edges may use different coordinates. Theorem 2 gives the contraction.

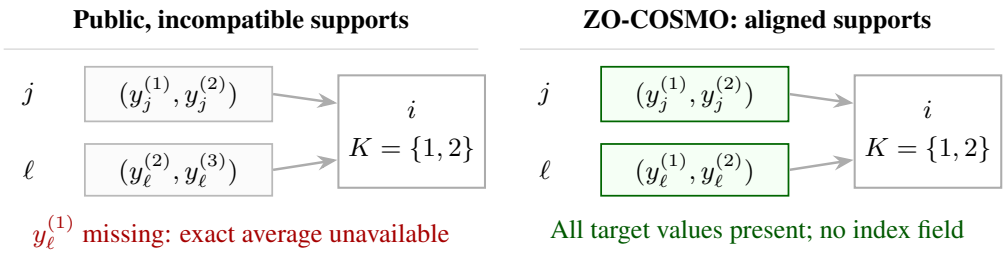


Figure 1: Both protocols know the supports. Only the right messages suffice for exact mixing on $K = \{1, 2\}$.

Algorithm 1 ZO-COSMO Core Update

Input: W, η, μ, q and public random tape $\{R_t\}_{t<T}$; initialize $x_{i,0}$.
for $t = 0, \ldots, T-1$ **do**
  From R_t, reconstruct $\mathcal{S}_t$ and signs ϵ_t; set $P_t = \text{diag}(\mathbf{1}_{\mathcal{S}_t})$ and $u_t = P_t\epsilon_t$.
  for all agents i in parallel **do**
    $\hat{g}_{i,t} = \frac{d}{q}\frac{f_i(x_{i,t}+\mu u_t;\xi_{i,t}) - f_i(x_{i,t}-\mu u_t;\xi_{i,t})}{2\mu} u_t$.
    $y_{i,t} = x_{i,t} - \eta\hat{g}_{i,t}$.
    Broadcast ordered values $\{y_{i,t}^{(j)} : j \in \mathcal{S}_t\}$, with no coordinate indices.
  end for
  for all agents i in parallel **do**
    $x_{i,t+1} = (I - P_t)y_{i,t} + P_t \sum_{\ell=1}^{N} w_{i\ell} y_{\ell,t}$.
  end for
end for
Output: $x_{i,T}$ or the network average $\bar{x}_T$.

The sparse-momentum variant changes only the local direction in Algorithm 1. It replaces $\hat{g}_{i,t}$ by $P_t m_{i,t+1}$, where

$$m_{i,t+1} = (I - P_t)m_{i,t} + P_t\big(\beta m_{i,t} + (1-\beta)\hat{g}_{i,t}\big). \tag{3}$$

The public support, masked mixing rule, and transmitted message are unchanged. Theorem 4 gives the sharper core rate; Appendix Theorem 5 proves convergence with sparse momentum for global and perfect-matching updates. Experiments report β explicitly.

4 THEORETICAL ANALYSIS

The analysis proceeds from decodability to mixing efficiency and then optimization. Exactness refers to coordinate recovery and the ideal arithmetic update; B_{val} accounts for a fixed-width scalar representation, without a separate quantization-error guarantee.

4.1 NOTATION

Let $F_i(x) = \mathbb{E}_{\xi_i}[f_i(x;\xi_i)]$, $F = N^{-1}\sum_i F_i$, and $\bar{x}_t = N^{-1}\sum_i x_{i,t}$. Before round t, an independent public coin R_t draws a uniform q-subset $\mathcal{S}_t$ and independent Rademacher signs. Define $P_t = \mathrm{diag}(\mathbf{1}_{\mathcal{S}_t})$ and $u_t = P_t\epsilon_t$. Then $\|u_t\|^2 = q$ and

$$\mathbb{E}[u_t u_t^\top] = (q/d)I_d. \tag{4}$$

The local two-query estimator is

$$\hat{g}_{i,t} = \frac{d}{q}\frac{f_i(x_{i,t}+\mu u_t;\xi_{i,t}) - f_i(x_{i,t}-\mu u_t;\xi_{i,t})}{2\mu}u_t. \tag{5}$$

For an ordered support $S = \{s_1 < \cdots < s_q\}$, define

$$\mathsf{Res}_S z = (z_{s_1},\ldots,z_{s_q}), \qquad \mathsf{Lift}_S(v) = \sum_{k=1}^{q} v_k e_{s_k}. \tag{6}$$

Then $\mathsf{Res}_S\mathsf{Lift}_S = I_q$, $\mathsf{Lift}_S\mathsf{Res}_S = P_S$, and $\|\mathsf{Lift}_S(v)\| = \|v\|$. In the protocol, a sender transmits $v = \mathsf{Res}_{S_t} z$ and the receiver reconstructs $\mathsf{Lift}_{S_t}(v)$ from the public support.

4.2 ASSUMPTIONS

Assumption 1. *Each F_i is L-smooth, and $F(x) \geq F^* > -\infty$ for all x.*

Assumption 2. *The mixing matrix W is nonnegative, symmetric, and doubly stochastic, with $\rho = \|W - N^{-1}\mathbf{1}\mathbf{1}^\top\|_2 < 1$.*

Assumption 3. *For each node i, $\mathbb{E}_{\xi_i}[\nabla f_i(x;\xi_i)] = \nabla F_i(x)$, $\mathbb{E}_{\xi_i}\|\nabla f_i(x;\xi_i) - \nabla F_i(x)\|^2 \leq \sigma^2$, and along the iterates $\mathbb{E}_{\xi_i}\|\nabla f_i(x;\xi_i)\|^2 \leq G^2 + \sigma^2$. At every round, the oracle samples are fresh and conditionally independent across nodes given the iterates and public direction.*

Assumption 4. *For every node i, point x, oracle sample ξ, and sparse direction u used by the algorithm, $|f_i(x+\mu u;\xi) - f_i(x-\mu u;\xi) - 2\mu\nabla f_i(x;\xi)^\top u| \leq (L_3/3)\mu^3\|u\|^3$.*

For later use, define the finite-difference bias scale and the per-node estimator moment bound as

$$B_\mu = \frac{L_3}{6}\mu^2 dq, \qquad M^2_{\mathrm{node}} = 2d(G^2+\sigma^2) + 2B_\mu^2. \tag{7}$$

4.3 COMMUNICATION ACCOUNTING

Definition 1. *A one-hop q-sparse current-state message from j to i contains ordered values $\mathsf{Res}_{S_{ji}}(y_j)$. The decoder knows its own state, W, and the public transcript, but receives no other part of y_j and does not replay earlier updates. In a private-support protocol the sender encodes S_{ji}; in a public-coin protocol both endpoints reconstruct it before communication.*

Theorem 1. *Fix receiver i, target support $K \subset [d]$, and message supports S_{ji} independently of the peer states, with $P_K = \mathrm{diag}(\mathbf{1}_K)$. From one ordered message $\mathsf{Res}_{S_{ji}}(y_j)$ per remote neighbor, a decoder can recover*

$$P_K \sum_{j=1}^{N} w_{ij} y_j \tag{8}$$

for every collection of peer states if and only if $K \subseteq S_{ji}$ for every $j \neq i$ with $w_{ij} > 0$. If $|K| = |S_{ji}| = q$, this requires $S_{ji} = K$. Moreover, if $k \in K \setminus S_{ji}$ and $\|y_j\|_\infty \leq R$, every decoder has worst-case coordinate error at least $w_{ij}R$.

Proof. See Appendix A.5. □

A zero-error support field costs at least $H(S_t \mid R_t)$ bits in expectation (Appendix Proposition 2). Public supports have zero conditional entropy; a private uniform q-subset costs $\log_2\binom{d}{q}$ bits. Reusing a data-dependent support over K rounds amortizes its description to $H(S \mid R)/K$, but fixes those coordinates throughout the block.

Consequently, with B_{val} value bits, each directed ZO-COSMO link transmits qB_{val} bits. An arbitrary worst-case support, or a uniform private support in expectation, needs $qB_{\mathrm{val}} + \log_2 \binom{d}{q}$ bits, while coordinate-list sparse encoding sends

$$q(B_{\mathrm{val}} + \lceil \log_2 d \rceil). \tag{9}$$

The binomial bounds in Appendix A.2 quantify the private-support cost. Reusing a full-model Qwen mask for ten rounds reduces coordinate-list overhead from $2.03\times$ to $1.10\times$, with the same coordinates used throughout the block. Public node-specific seeds also regenerate independent random supports without recurring metadata; we include this encoding control in the experiments.

4.4 Edge-Local Mixing

Theorem 1 applies locally: supports must agree across an active mixing edge, but disjoint edges need not share a support. Let M_t be a public matching, $L_e = (e_i - e_j)(e_i - e_j)^\top$ for $e = \{i, j\}$, and let each $P_{e,t}$ be an independent uniform rank-q coordinate projector. The support-labeled mixing step is

$$X_{t+1} = Y_t - \frac{1}{2} \sum_{e \in M_t} L_e Y_t P_{e,t}. \tag{10}$$

Theorem 2. *Let $\Pi = I - N^{-1}\mathbf{1}\mathbf{1}^\top$ and $\bar{L} = \mathbb{E}[\sum_{e \in M_t} L_e]$. For any fixed input Y_t independent of the current matching and supports,* (10) *preserves the network average and satisfies*

$$\mathbb{E}_P \|\Pi X_{t+1}\|_F^2 = \|\Pi Y_t\|_F^2 - \frac{q}{2d} \sum_{\{i,j\} \in M_t} \|y_{i,t} - y_{j,t}\|^2, \tag{11}$$

$$\mathbb{E}_{M,P} \|\Pi X_{t+1}\|_F^2 \le \left(1 - \frac{q\lambda_2(\bar{L})}{2d}\right) \|\Pi Y_t\|_F^2. \tag{12}$$

The identities require uniform marginal supports, not independence across disjoint edges. Each active endpoint sends qB_{val} bits and no indices.

Proof. See Appendix A.7. □

The optimization update Y_t shares the current query mask with mixing. Appendix A.20 retains this dependence by applying contraction to the past state and controlling the current update separately.

To compare matching schedules independently of a particular state, let

$$\bar{L}_k = \mathbb{E}\left[\sum_{e \in M_t} \mathbf{1}\{k \in S_{e,t}\} L_e\right], \qquad C_{\mathrm{dir}} = 2\mathbb{E}\left[\sum_{e \in M_t} |S_{e,t}|\right], \tag{13}$$

where C_{dir} is the expected number of directed scalar values sent in one round.

Theorem 3. *For any matching-based exact one-hop operator of the form* (10)*, allowing arbitrary state-independent public coordinate supports, the largest γ such that*

$$\mathbb{E}\|\Pi X_{t+1}\|_F^2 \le (1 - \gamma)\|\Pi Y_t\|_F^2 \tag{14}$$

for every fixed input Y_t independent of the sampled operator is

$$\gamma_\star = \frac{1}{2} \min_{k \in [d]} \lambda_2(\bar{L}_k) \le \frac{C_{\mathrm{dir}}}{2d(N-1)}. \tag{15}$$

With B_{val}-bit scalars, $B_{\mathrm{net}} \ge 2B_{\mathrm{val}} d(N-1)\gamma_\star$. For even N, uniform rank-q supports on uniformly random perfect matchings of K_N attain equality with $C_{\mathrm{dir}} = Nq$, using either a common support or independent edge supports.

Proof. See Appendix A.8. □

Equality requires the same isotropic expected mixing on every coordinate (Appendix A.8). Uniform perfect matchings on complete graphs satisfy this condition with either common or independent edge supports. Both attain the pure-mixing bound; Section 4.6 analyzes their different query statistics.

Appendix Proposition 3 gives the estimator bias B_μ and moment bound M_{node}^2 used below; a coordinate is refreshed after d/q rounds in expectation.

4.5 Convergence Bound

Let

$$\mathcal{E}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|x_{i,t} - \bar{x}_t\|^2 \tag{16}$$

be the parameter consensus error.

We use the conservative stability condition

$$0 < \eta \le \min\{L^{-1},\, c(Ld)^{-1}\}, \tag{17}$$

for a universal constant $c > 0$, and the canonical smoothing scale

$$\mu = \Theta((dq)^{-1/2}T^{-1/4}), \tag{18}$$

which balances the displayed finite-difference bias terms in the proof.

Theorem 4. *Let $\Delta_0 = F(\bar{x}_0) - F^*$ and suppose Assumptions 1–4 hold. Run Algorithm 1 from a common initialization with $\mu = \Theta((dq)^{-1/2}T^{-1/4})$ and $\eta = c_\eta T^{-1/2}$. For any fixed $c_\eta > 0$ and every horizon T large enough that* (17) *holds,*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \le \mathcal{O}\left(\frac{\Delta_0}{\sqrt{T}} + \frac{d\sigma^2}{N\sqrt{T}} + \frac{1}{T} + \frac{d^2(G^2+\sigma^2)}{q(1-\rho^2)^2T}\right). \tag{19}$$

Proof. The descent proof is in Appendix A.16; the budget algebra is in Appendix C.7. □

Remark 1. *The leading terms in* (19) *describe stochastic ZO descent; the last accounts for sparse mixing. Under a directed-link budget $\mathcal{B}$, $T = \lfloor \mathcal{B}/(qB_{\mathrm{val}})\rfloor$ converts the network term to $\mathcal{O}(d^2(G^2+\sigma^2)B_{\mathrm{val}}/((1-\rho^2)^2\mathcal{B}))$: its explicit $1/q$ factor cancels, while the spectral penalty remains. Smaller q buys more rounds and queries. Appendix C.7 gives the full bound, and Appendix A.15 proves the unavoidable q/d one-step mixing slowdown.*

For the momentum update, set $a_\beta = \beta/(1-\beta)$ and $\bar{m}_t = N^{-1}\sum_i m_{i,t}$. Average preservation gives the exact auxiliary recursion

$$z_t = \bar{x}_t - \eta a_\beta \bar{m}_t, \qquad z_{t+1} = z_t - \eta\bar{g}_t. \tag{20}$$

Appendix Theorem 5 bounds the distance from z_t to $\bar{x}_t$ and handles the shared query–mixing mask, proving convergence for both global and perfect-matching momentum updates.

4.6 Direction Correlation and Heterogeneity

Support agreement makes a message decodable; direction agreement also changes the update statistics. Hold the matching fixed, with $N = 2m$. Let $a_e = (\nabla F_i(x_i) + \nabla F_j(x_j))/2$ on pair $e = \{i,j\}$, $\bar{a} = m^{-1}\sum_e a_e$, and $H_M^2 = m^{-1}\sum_e \|a_e - \bar{a}\|^2$. The linearized two-query update is $v_e = D_e a_e$, where $D_e = (d/q)u_e u_e^\top$ and $\bar{v} = m^{-1}\sum_e v_e$.

Both endpoints share u_e. Across pairs, G shares the support and signs, S shares only the support, and I draws both independently. All three use the same matching and send q values per node.

Proposition 1. *Condition on the states and matching. With $V_s = \mathbb{E}\|\bar{v} - \bar{a}\|^2$ and $U_s = m^{-1}\sum_e \mathbb{E}\|v_e - \bar{v}\|^2$,*

$$V_{\mathrm{G}} = (d-1)\|\bar{a}\|^2, \qquad V_{\mathrm{I}} = \frac{d-1}{m}(\|\bar{a}\|^2 + H_M^2), \tag{21}$$

$$V_s + U_s = (d-1)\|\bar{a}\|^2 + dH_M^2, \qquad s \in \{\mathrm{G},\mathrm{S},\mathrm{I}\}. \tag{22}$$

Proof. See Appendix B.2, including the formula for S. □

Under G, averaging commutes with the random map: $\bar{v} = D\bar{a}$. Local gradient components that cancel in $\bar{a}$ therefore cancel before estimation. Independent maps trade this cancellation for averaging direction noise across pairs. For $d > 1$, I has smaller V_s precisely when $H_M^2 < (m-1)\|\bar{a}\|^2$; G benefits on the other side of this threshold. Equation (22) tracks the accompanying disagreement

injection. Correlation therefore changes the update itself, even when the schedule and payload are fixed.

With unequal quadratic Hessians, state disagreement controls the residual (Appendix Proposition 6). Let $\mathcal{A}$ denote masked pair averaging and V the applied update matrix. The two rows of V on each pair agree, giving $\mathcal{A}V = V$ even though queries and mixing share a mask. Appendix Proposition 7 extends the moments to oracle noise, stored momentum, and finite differences. Theorems 6 and 5 give, respectively, a sharper deterministic core rate for iid matchings and convergence with stochastic sparse momentum.

5 Experiments

We charge value and index fields per directed link. Topology comparisons fix per-node bits because the active edge sets differ; total network traffic is also recorded. Dense and refresh methods pay for every full-state exchange. Seed setup is separate from recurring payloads.

The controls isolate encoding, mixing, and local optimization. Explicit-index and seed-aware encodings measure the value of removing recurring metadata. Dense ZO-DSGD provides a dense-update reference; Com-DSZO tests adaptive compression. ZO-MARINA-type uses periodic dense refreshes rather than MARINA's gradient-difference estimator (Appendix D).

5.1 Synthetic Experiments

On 20-dimensional Rosenbrock with $N = 10$, the indexed control uses identical ZO-COSMO updates but redundantly sends each support. Com-DSZO (Hua et al., 2026) combines dense two-point updates with Top-k innovation compression and error-compensated consensus. All methods use two function evaluations per node and round.

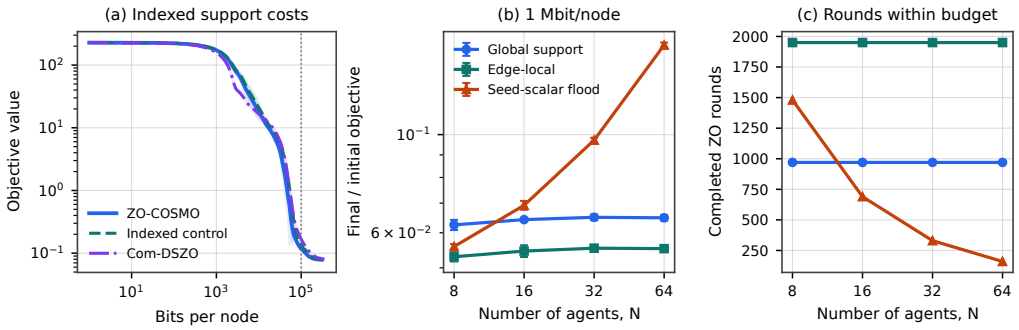


Figure 2: Communication-limited Rosenbrock. Bands and bars show one standard deviation over five seeds.

Table 1: Rosenbrock at 0.1M bits per node with $q = 1$ (five seeds). Lower is better.

Method	Ring	Erdős-Rényi	Grid
ZO-COSMO (Ours)	$\mathbf{0.083 \pm 0.006}$	$\mathbf{0.119 \pm 0.011}$	$\mathbf{0.093 \pm 0.006}$
Same updates + indices	0.088 ± 0.006	0.136 ± 0.015	0.103 ± 0.006
Com-DSZO (Top-k)	0.097 ± 0.007	0.171 ± 0.018	0.115 ± 0.007

At the marked 0.1M-bit budget, removing repeated indices buys 15–17% more rounds. ZO-COSMO has the lowest objective in all 15 paired seed–topology comparisons. At 0.3M bits the methods have nearly converged, so the gap shrinks below 0.004 on the Erdős-Rényi graph. The identical-update control attributes this gain to the extra optimization rounds purchased by value-only messages.

Panels (b)–(c) compare one-hop global support, edge-local support, and an optimistic seed–scalar flood on rings. At 1 Mbit per node, edge-local ZO-COSMO completes 1950 logged rounds, versus 970 global-support rounds, and has the lowest objective for every $N \in \{8, 16, 32, 64\}$. Flooding is competitive at $N = 8$, but its all-node dissemination cost leaves only 160 rounds at $N = 64$, where its relative objective is 0.159 ± 0.002 versus 0.055 ± 0.0004 for edge-local mixing. The separate topology stress test reaches 64 agents on complete, grid, and ring graphs (Appendix Figure 14).

5.2 Mechanism Experiment

We compare G, S, and I with $N = 8$, $d = 32$, $q = 4$, and paired matchings. Local quadratics have unequal positive diagonal Hessians and centered linear terms with $N^{-1}\sum_i \|b_i\|^2 = 4$. Noisy queries add fresh independent $\epsilon_i^\top x$, with $\mathbb{E}\|\epsilon_i\|^2 = 0.25$, reusing each sample within its query pair.

Figure 3 uses iid ring matchings and $\eta = 0.004$. With oracle noise, G reaches $(1.06 \pm 0.30) \times 10^{-3}$ versus $(9.60 \pm 1.69) \times 10^{-3}$ for I; with sparse momentum $\beta = 0.9$ at the same step, the corresponding values are $(1.00 \pm 0.13) \times 10^{-3}$ and $(10.60 \pm 2.70) \times 10^{-3}$. Noise leaves a nonzero floor.

Matching the first update by increasing the momentum step to 0.04 reverses the comparison: G ends at 1.30 ± 1.42, versus 0.131 ± 0.021 for I. Appendix B.10 includes this control, zero-linear-heterogeneity cases, and all 1536 runs across two steps and topologies.

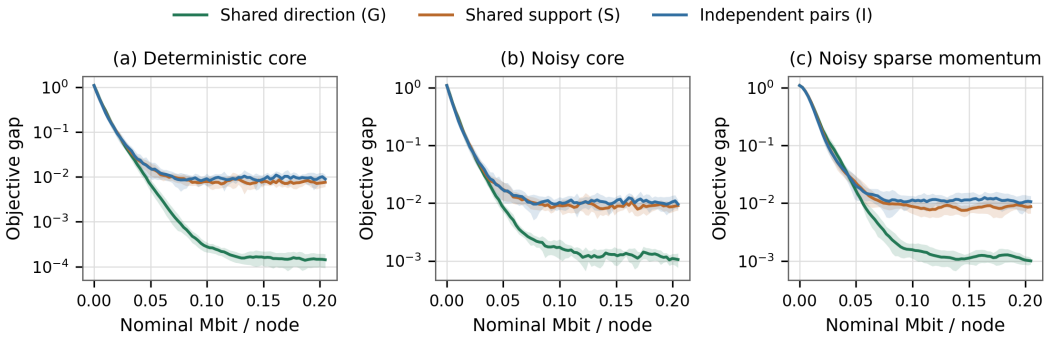


Figure 3: Different local curvatures on an eight-node ring. Eight-seed mean ± std at equal queries and nominal payload; (c) uses the same step as (b).

5.3 Zeroth-Order Fine-Tuning for Large Language Models

We study Qwen2-7B (Yang et al., 2024) on QNLI with 2.52M LoRA coordinates (Hu et al., 2022). Workers maintain separate float32 adapter and momentum states on disjoint data shards; complete and ring overlays use $N \in \{2, 4, 8\}$ logical workers. QNLI was selected before training by a fixed task-screening rule (Appendix D). We evaluate all 5463 validation examples.

Value-only communication extends the affordable training horizon. Each explicit coordinate costs 22 bits in addition to its 32-bit value; at $B = 26.214$ Mbit per directed link, ZO-COSMO completes 100 rounds versus 59 for explicit-index Rand-k. Over ten paired seeds, QNLI accuracy increases from $61.75 \pm 3.41\%$ to $65.40 \pm 3.57\%$: a 3.65-point gain (95% bootstrap CI $[0.49, 6.86]$). Mean query loss decreases by 0.182 (CI $[0.093, 0.273]$). Figure 4 compares the resulting models at the same communication budget.

Node-specific public seeds also permit value-only encoding of independent Rand-k. This post-hoc lossless re-encoding retains its 100-round accuracy of $66.48 \pm 7.36\%$ at the same budget. ZO-COSMO's paired accuracy difference is -1.08 points (CI $[-5.85, 3.67]$), with mean query loss lower by 0.091 (CI $[0.006, 0.175]$). The control separates the benefit of removing indices from that of coordinating queries and peer-state mixing.

Sparse momentum provides a separate optimization gain. With support, communication, and first-update scale $(1-\beta)\eta$ matched, it raises accuracy from $61.48 \pm 3.31\%$ to $65.40 \pm 3.57\%$ (paired $+3.92$ points, CI $[1.67, 6.08]$). Matching the initial scale isolates the contribution of stored momentum; Appendix Table 15 reports both step-size controls.

The representation saving grows with model dimension. In the single-worker full-model study, sparse updates over 6.53B trainable coordinates reach $95.03 \pm 0.13\%$ SST-2 accuracy. Encoding these same updates takes 262.1K bits per step with public support, compared with 532.5K bits using explicit indices, a $2.03\times$ reduction.

Edge-local mixing extends the budget advantage to larger networks. At 26.214 Mbit per node, matching permits 100 rounds, versus 14 all-neighbor rounds on the eight-worker complete graph and 50 on the ring. Five-seed accuracy rises from $61.44 \pm 0.24\%$ to $64.85 \pm 3.03\%$ on the complete graph

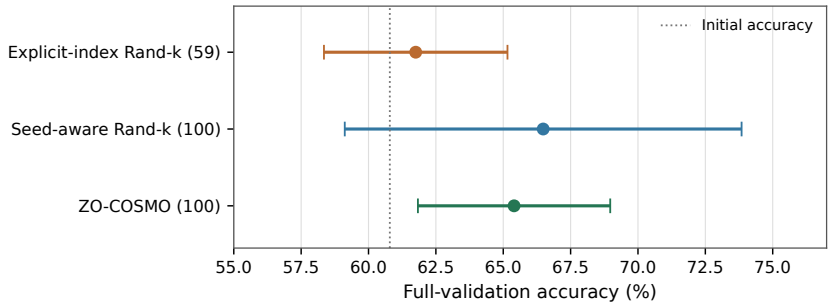


Figure 4: QNLI at 26.214 Mbit per link (ten-seed mean ± std). Seed-aware Rand-k is a lossless re-encoding of existing runs.

and from $61.09 \pm 0.63\%$ to $63.62 \pm 0.85\%$ on the ring: gains of 3.42 points (CI $[1.06, 5.63]$) and 2.53 points (CI $[1.87, 3.07]$). With query correlation fixed on the four-worker ring, matching lowers same-budget label NLL by 0.057 (CI $[0.017, 0.105]$). Figure 19 includes all four topology–size settings and equal-round controls.

Long-horizon training remains effective with a common tenfold step reduction after round 100. At 1540 rounds on the four-worker ring, the three couplings reach 74.07–77.82% accuracy over five seeds. With matching and support draws fixed, independent edge signs improve accuracy by 3.75 points (CI $[0.24, 6.79]$) and lower label NLL by 0.027 (CI $[0.009, 0.049]$). The gain comes from query correlation at unchanged scheduling and support selection (Appendix D.3).

6 Conclusion

ZO-COSMO performs decentralized two-query updates with q values per active link and no recurring coordinate indices. The analysis identifies the support compatibility needed for one-hop mixing, quantifies contraction per communicated scalar, and explains how direction correlation trades estimation error against disagreement. Synthetic and Qwen experiments show fixed-budget gains from compact messages and matching schedules, together with separate benefits from sparse momentum and edge-sign independence.

Reproducibility Statement

The supplement and code provide the proofs, payload formulas, baseline definitions, hyperparameters, seeds, hardware, fixed-budget protocol, and raw CSVs underlying every reported experimental aggregate.

AI Use Statement

Generative AI tools assisted with literature discovery for finding the related work; editing and polishing the manuscript. The authors reviewed the AI-assisted material and take full responsibility for the final content.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pp. 14668–14679, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569, 2018.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

HanQin Cai, Daniel Mckenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (ZORO): Approximately sparse gradients and adaptive sampling. *arXiv preprint arXiv:2003.13001*, 2020.

Zhijie Cai, Haolong Chen, Guangxu Zhu, Qingjiang Shi, and Kaibin Huang. Feedsign: Robust full-parameter federated fine-tuning of large models with extremely low communication overhead of one bit. *IEEE Transactions on Mobile Computing*, 25(9):14225–14238, 2026. doi: 10.1109/TMC.2026.3677600.

Linjing Chen, Antai Xie, Xinlei Yi, Xiaoqiang Ren, and Xiaofan Wang. Compressed momentum-based single-point zero-order algorithm for stochastic distributed nonconvex optimization. *arXiv preprint arXiv:2512.06366*, 2025. URL https://arxiv.org/abs/2512.06366.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. ZO-AdaMM: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, pp. 7204–7215, 2019.

Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. SIAM Philadelphia, 2009.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, et al. Gradientless descent: High-dimensional zeroth-order optimization. *arXiv preprint arXiv:1911.06317*, 2019.

Eduard Gorbunov, Konstantin P. Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster nonconvex distributed learning with compression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3788–3798. PMLR, 2021.

Davood Hajinezhad, Mingyi Hong, and Alfredo Garcia. ZONE: Zeroth-order nonconvex multiagent optimization over networks. *IEEE Transactions on Automatic Control*, 64(10):3995–4010, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Youqing Hua, Shuai Liu, Yiguang Hong, and Wei Ren. Distributed stochastic zeroth-order optimization with compressed communication. *IEEE Transactions on Automatic Control*, 71(2):1294–1301, 2026. doi: 10.1109/TAC.2025.3610109.

Jihun Kim and Namhoon Lee. Seedflood: A step toward scalable decentralized training of LLMs. *arXiv preprint arXiv:2602.18181*, 2026. URL https://arxiv.org/abs/2602.18181.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pp. 3478–3487, 2019.

Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations (ICLR)*, 2020.

David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *arXiv preprint arXiv:2003.02684*, 2020.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1–2):237–284, 2020. doi: 10.1007/s10107-018-1355-4.

Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Achieving dimension-free communication in federated learning via zeroth-order optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=omrLHFzC37.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

Liu Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimization via variance reduction method. *arXiv preprint arXiv:1805.11811*, 2018a.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3727–3737, 2018b.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075, 2023.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Yide Ran, Wentao Guo, Jingwei Sun, Yanzhou Pan, Xiaodong Yu, Hao Wang, Jianwen Xie, Yiran Chen, Denghui Zhang, and Zhaozhuo Xu. Mitigating non-IID drift in zeroth-order federated LLM fine-tuning with transferable sparsity. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2DuMBKVbX2.

Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic, and Soummya Kar. Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach. In *IEEE Conference on Decision and Control*, pp. 4951–4958, 2018.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Shaohuai Shi, Qiang Wang, Xiaowen Chu, Bo Li, Yang Qin, Ruihao Liu, and Xinxiao Zhao. Communication-efficient distributed deep learning with merged gradient sparsification on gpus. In *IEEE INFOCOM 2020 - IEEE International Conference on Computer Communications*, pp. 406–415. IEEE, 2020.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GugZ5DzzAu.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems*, pp. 7652–7662, 2018.

Yujie Tang and Na Li. Distributed zero-order algorithms for nonconvex multi-agent optimization. In *Annual Allerton Conference on Communication, Control, and Computing*, pp. 781–786, 2019.

A. S. Veprikov, E. D. Petrov, G. V. Evseev, and A. N. Beznosikov. Zero order algorithm for decentralized optimization problems. *Doklady Mathematics*, 110(Suppl 1):S261–S277, 2024. doi: 10.1134/S1064562424602336.

Haonan Wang, Xinlei Yi, and Yiguang Hong. Compressed zeroth-order algorithm for stochastic distributed nonconvex optimization. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 4364–4369, 2025. doi: 10.1109/CDC57313.2025.11312521.

Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1356–1365, 2018.

An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4475–4485, 2025.

Deming Yuan and Daniel WC Ho. Randomized gradient-free method for multiagent optimization over time-varying networks. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6): 1342–1347, 2015. doi: 10.1109/TNNLS.2014.2339035.

Deming Yuan, Shengyuan Xu, and Junwei Lu. Gradient-free method for distributed multi-agent optimization via push-sum algorithms. *International Journal of Robust and Nonlinear Control*, 25 (10):1569–1580, 2015. doi: 10.1002/rnc.3164.

Shengjun Zhang and Colleen P Bailey. Accelerated zeroth-order algorithm for stochastic distributed non-convex optimization. In *2022 American Control Conference (ACC)*, pp. 4274–4279, 2022. doi: 10.23919/ACC53348.2022.9867306.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 2024.

A Proofs

We give the details behind Section 4. Constants denoted by C are numerical constants independent of T, d, q, N, η, and μ.

A.1 Proof Structure

We first establish the support-coding and decoding results. The optimization proof then controls the finite-difference estimate along u_t and disagreement after mixing on the same support, retaining their dependence in the conditional expectations.

Smoothness at the network average gives

$$F(\bar{x}_{t+1}) \le F(\bar{x}_t) - \eta\langle \nabla F(\bar{x}_t), h_t\rangle + \frac{L\eta^2}{2}\|h_t\|^2, \tag{23}$$

For Algorithm 1, $h_t = \bar{g}_t$, and the direction decomposes as

$$h_t = \nabla F(\bar{x}_t) + \underbrace{(\bar{g}_t - \mathbb{E}_t \bar{g}_t)}_{\text{oracle and direction noise}} + \underbrace{(\mathbb{E}_t \bar{g}_t - \nabla F(\bar{x}_t))}_{\text{finite-difference and disagreement bias}}. \tag{24}$$

The estimator lemmas control the two error terms in (24). The masked disagreement recursion gives the price of mixing a coordinate only when it belongs to the public support. Substitution into (23), followed by telescoping, proves Theorem 4.

The filtration is fixed as follows. Let $\mathcal{F}_t$ contain the iterates, algorithm state, graph, past public coins $R_0, \dots, R_{t-1}$, and oracle samples through round $t-1$. The fresh public coin R_t, and hence the support and signs at round t, is independent of $\mathcal{F}_t$. Conditional on $\mathcal{F}_t$ and R_t, the oracle variables $\xi_{i,t}$ are independent across nodes. Worker averaging reduces independent oracle noise, while shared direction randomness remains in the conditional moments. Appendix A.4 specifies the seeded implementation and the role of the ideal public-coin assumption.

We write $\mathbb{E}_S[\cdot]$ for expectation over the support, $\mathbb{E}_u[\cdot]$ for expectation over support and signs, and $\mathbb{E}_\xi[\cdot]$ for expectation over oracle samples. An omitted subscript includes all randomness. Here P_t projects onto $\mathcal{S}_t$, u_t is the sparse Rademacher direction, and ρ is the spectral radius of $W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. Public information determines the support at both endpoints; its random selection still contributes to the estimator variance.

A.2 Conditional Support Cost

Proposition 2. *Any zero-error uniquely decodable support field needs at least $H(S_t \mid R_t)$ bits in expectation; with B_{val} value bits, the total is at least*

$$qB_{\mathrm{val}} + H(S_t \mid R_t). \tag{25}$$

If one data-dependent support S is announced and reused for a block of K rounds, its amortized per-round cost is at least

$$qB_{\mathrm{val}} + \frac{H(S \mid R)}{K}. \tag{26}$$

Zero expected support length requires the support to be determined by public information. An unreconstructable uniform q-subset costs $\log_2 \binom{d}{q}$ bits, of order $q\log(d/q)$ for $q \le d/2$. Public supports remove this field; reused adaptive masks amortize it as in (26)*.*

The elementary binomial estimates used in the main text are

$$q\log_2(d/q) \le \log_2 \binom{d}{q} \le q\log_2(ed/q). \tag{27}$$

Fix a communication round and condition on the public information $R_t = r$. A zero-error decoder must recover the realized support S_t from the transmitted support codeword and r. Thus, for each r, the code must be uniquely decodable for the conditional source $S_t \mid R_t = r$. The source-coding converse gives

$$\mathbb{E}[\ell_t \mid R_t = r] \ge H(S_t \mid R_t = r), \tag{28}$$

where ℓ_t is the number of transmitted support bits in that round. Taking expectation over R_t yields

$$\mathbb{E}\ell_t \ge \mathbb{E}_{R_t} H(S_t \mid R_t) = H(S_t \mid R_t). \tag{29}$$

The value payload contributes qB_{val} additional bits, giving (25). Because codeword lengths are nonnegative, zero expected support length forces $H(S_t \mid R_t) = 0$. For a discrete support this is equivalent to S_t being a function of R_t almost surely. If $S_t = \Phi(R_t, t)$, this condition holds. If instead S_t is uniform over $\binom{[d]}{q}$ and independent of R_t, then $H(S_t \mid R_t) = H(S_t) = \log_2 \binom{d}{q}$; Stirling's bounds give $\Theta(q\log(d/q))$ when $1 \le q \le d/2$.

For a block-reused support, condition on the public information available when the block begins. Zero-error recovery of the private support still requires expected length at least $H(S \mid R)$. The block

then carries KqB_{val} value bits and at least $H(S \mid R)$ support bits. Dividing by K proves (26). This is a lower bound on describing a chosen mask; any communication used to agree on a data-dependent mask across decentralized nodes is additional.

A.3 Private-Support Code Length

Let

$$\mathcal{U}_{d,q} := \{S \subset [d] : |S| = q\}, \qquad |\mathcal{U}_{d,q}| = \binom{d}{q}. \tag{30}$$

For a fixed-length zero-error code of length L, distinct supports require distinct codewords, so $\binom{d}{q} \le 2^L$. For a prefix-free code whose maximum length is L, Kraft's inequality gives

$$1 \ge \sum_{S \in \mathcal{U}_{d,q}} 2^{-\ell(S)} \ge \binom{d}{q} 2^{-L}. \tag{31}$$

Either model therefore requires

$$\lceil \log_2 |\mathcal{U}_{d,q}| \rceil = \left\lceil \log_2 \binom{d}{q} \right\rceil \tag{32}$$

bits in the worst case.

For the expected-length statement, assume S is uniform over $\mathcal{U}_{d,q}$. Any uniquely decodable zero-error code has expected length at least the entropy of the source, so

$$\mathbb{E}[\ell(S)] \ge H(S) = \log_2 |\mathcal{U}_{d,q}| = \log_2 \binom{d}{q}. \tag{33}$$

The standard source-coding converse gives the claim.

The displayed $\Theta(q \log(d/q))$ order follows from the standard binomial estimates

$$\left(\frac{d}{q}\right)^q \le \binom{d}{q} \le \left(\frac{ed}{q}\right)^q. \tag{34}$$

Taking base-two logarithms gives (27) for $q \le d/2$. Coordinate lists cost $q\lceil \log_2 d \rceil$ bits; optimal private-support codes may be shorter. Publicly reconstructible supports require zero support bits. These are the three encodings compared in the communication analysis.

In the public-coin model used by ZO-COSMO, the support is generated as

$$S_t = \Phi(R_t, t), \tag{35}$$

where R_t is known to both sender and receiver and Φ is a deterministic rule. Therefore S_t is measurable with respect to R_t, and

$$H(S_t \mid R_t) = 0. \tag{36}$$

Thus no support field is needed. This coding observation applies to any publicly reconstructible support, including independent Rand-k supports generated from node-specific seeds.

A.4 Public-Support Lifting

The implementation reconstructs each round from the master seed s_0 and public identifiers. In the experiment code, BLAKE2b folds these identifiers into a NumPy generator seed:

$$r_t = \mathrm{Hash}(s_0, t), \qquad (S_t, \varepsilon_t) = \mathrm{SampleSparseDirection}(r_t, d, q), \tag{37}$$

where $S_t \subset [d]$ has cardinality q and $\varepsilon_t \in \{\pm 1\}^{S_t}$. The sparse direction is $u_{t,j} = \varepsilon_{t,j} \mathbf{1}\{j \in S_t\}$. Node identifiers or canonically ordered edge endpoints select local streams when required. With the same generator, ordering convention, and public identifiers, endpoints reconstruct the same support and signs directly.

The communicated message is the ordered value vector

$$m_{i,t} = \big(v_{i,t,1}, \ldots, v_{i,t,q}\big), \tag{38}$$

where $v_{i,t,k}$ is attached to the k-th coordinate in the deterministic ordering of S_t. No coordinate identifier appears in (38). If $S_t = \{s_{t,1} < \cdots < s_{t,q}\}$, then decoding is exactly the public-support lift

$$\mathrm{Decode}(m_{i,t}; s_0, t) = \mathsf{Lift}_{S_t}(m_{i,t}) = \sum_{k=1}^{q} v_{i,t,k} e_{s_{t,k}}, \tag{39}$$

with all other coordinates equal to zero.

Decoding and average preservation hold pathwise for every support-compatible sequence, including seeded sequences. The contraction and convergence rates assume fresh uniform draws in the ideal public-coin model. The finite-seed implementation reproduces the message identities exactly; extending the stochastic rates to its pseudorandom generator would require a separate randomness approximation or cryptographic argument.

Endpoints share only the generator specification, seed, round counter, d, and q; data, losses, gradients, and momentum remain local. With a κ-bit master seed, initialization contributes $\kappa + O(\log d + \log q + \log T)$ bits per neighbor when charged explicitly, rather than a recurring support field.

Lemma 1. *Assume two neighboring agents use the same tuple (s_0, t, d, q) and the same deterministic sampling rule in* (37)*. Then the decoder* (39) *reconstructs the sender's sparse vector from ordered values alone. Consequently, exact sparse-message decoding requires no per-round coordinate metadata.*

Proof. Once the sampling rule is fixed, S_t and ε_t are deterministic functions of (s_0, t, d, q). Both endpoints therefore obtain the same ordered support. The map Lift_{S_t} places the k-th received value on the k-th public coordinate, so the decoded vector is the one formed by the sender. □

A.5 Consequences of One-Hop Support Compatibility

Proof of Theorem 1. If $K \subseteq S_{ji}$ for every positive-weight remote neighbor, receiver i lifts each ordered message, reads its coordinates in K, and adds the known self term. This proves sufficiency. For necessity, suppose $w_{ij} > 0$ and $k \in K \setminus S_{ji}$. Hold every state fixed except $y_j^{(k)}$, setting that entry to R in one instance and $-R$ in another. The receiver's state, public transcript, supports, and messages are identical, while the correct k-th outputs differ by $2w_{ij}R$. Any common output has error at least half this separation on one instance. Taking arbitrary $R > 0$ rules out zero-error recovery; under the stated bound it gives the minimax error $w_{ij}R$. If both supports have cardinality q, containment implies equality. □

Theorem 1 concerns exact current-state mixing. PermK instead coordinates compressors to control the error of an aggregate estimate (Szlendak et al., 2022); the inputs at different workers need not be identical. Complementary supports can serve that purpose without exposing every peer on a common target coordinate. Likewise, a seed–scalar pair reconstructs an update, but not an arbitrary state to which past updates have been applied. For randomized decoders, the same two indistinguishable states give an expected absolute-error lower bound: $|a - w_{ij}R| + |a + w_{ij}R| \geq 2w_{ij}R$ for every output a, and taking expectations preserves the inequality. Setting the unknown contribution to zero attains this coordinatewise bound when all other contributions are known.

The condition is local to an active communication group. If all neighbors of i participate in one Metropolis step, they need a common target support at i. If communication is scheduled as a matching, only the two endpoints of each active edge must agree. This observation leads to the edge-local method below.

A.6 Edge-Local ZO-COSMO

Algorithm 2 uses public perfect matchings. For a maximal matching, unmatched nodes retain their states. A positive contraction in Theorem 2 requires a connected expected union graph, equivalently $\lambda_2(\bar{L}) > 0$.

Every edge support is keyed by the unordered edge id, so its endpoints reconstruct the same ordered coordinates. Different matched edges may reconstruct different supports because they have disjoint

Algorithm 2 Edge-Local ZO-COSMO

Input: graph G, η, μ, q, and public tape; initialize $x_{i,0}$.
for $t = 0, \dots, T-1$ **do**
  Reconstruct a public matching M_t.
  for all $e = \{i, j\} \in M_t$ in parallel **do**
    From (R_t, e), reconstruct support $S_{e,t}$, signs $\epsilon_{e,t}$, and $u_{e,t} = P_{e,t}\epsilon_{e,t}$.
    for all $r \in \{i, j\}$ in parallel **do**
      $\hat{g}_{r,t} = \frac{d}{q}\frac{f_r(x_{r,t}+\mu u_{e,t};\xi_{r,t}) - f_r(x_{r,t}-\mu u_{e,t};\xi_{r,t})}{2\mu} u_{e,t}$; $y_{r,t} = x_{r,t} - \eta\hat{g}_{r,t}$.
      Send $\mathsf{Res}_{S_{e,t}}(y_{r,t})$ to the other endpoint.
    end for
    $x_{i,t+1} = (I - P_{e,t})y_{i,t} + \frac{1}{2}P_{e,t}(y_{i,t} + y_{j,t})$; update $x_{j,t+1}$ symmetrically.
  end for
end for

receivers. The code alternates the two perfect matchings of an even ring; on a complete graph it publicly permutes the nodes and pairs adjacent entries. The momentum variant replaces $\hat{g}_{r,t}$ in the local step by $P_{e,t}m_{r,t+1}$ using (3) with $P_t = P_{e,t}$.

A.7 Proof of Theorem 2

For $e = \{i, j\}$, write $Q_e(Z) = \frac{1}{2}L_e Z P_{e,t}$. Since $L_e^2 = 2L_e$ and $P_{e,t}^2 = P_{e,t}$, Q_e is an orthogonal projector under the Frobenius inner product. Distinct edges in a matching have disjoint endpoints, hence $L_e L_{e'} = 0$ and the projectors $Q_e, Q_{e'}$ are orthogonal even when their coordinate supports overlap. Thus $Q = \sum_{e \in M_t} Q_e$ is itself an orthogonal projector and (10) is $X_{t+1} = (I - Q)Y_t$.

Because $\mathbf{1}^\top L_e = 0$, the network average is unchanged. Moreover, QY_t lies in the disagreement subspace and $Q\Pi Y_t = QY_t$. Pythagoras therefore gives

$$\|\Pi X_{t+1}\|_F^2 = \|(I - Q)\Pi Y_t\|_F^2 \tag{40}$$

$$= \|\Pi Y_t\|_F^2 - \|QY_t\|_F^2 \tag{41}$$

$$= \|\Pi Y_t\|_F^2 - \frac{1}{2}\sum_{\{i,j\} \in M_t} \|P_{e,t}(y_{i,t} - y_{j,t})\|^2. \tag{42}$$

For a uniform q-coordinate projector and a fixed vector v independent of it, $\mathbb{E}\|P_{e,t}v\|^2 = (q/d)\|v\|^2$. Holding Y_t fixed while taking the support expectation in (42) proves (11). No cross-edge independence is used. Averaging also over the matching yields

$$\mathbb{E}_{M,P}\|\Pi X_{t+1}\|_F^2 = \|\Pi Y_t\|_F^2 - \frac{q}{2d}\operatorname{tr}(Y_t^\top \bar{L} Y_t). \tag{43}$$

Since $\bar{L}\mathbf{1} = 0$ and $\bar{L} \succeq 0$, the variational definition of λ_2 gives $\operatorname{tr}(Y_t^\top \bar{L} Y_t) \geq \lambda_2(\bar{L})\|\Pi Y_t\|_F^2$, proving (12).

For a uniformly random perfect matching of the complete graph, $\bar{L} = L_{K_N}/(N-1)$ and $\lambda_2(\bar{L}) = N/(N-1)$. Drawing either perfect-matching phase of an even ring uniformly gives $\bar{L} = L_{\text{ring}}/2$ and

$$\lambda_2(\bar{L}) = 1 - \cos(2\pi/N). \tag{44}$$

These constants concern independently sampled matching phases and a fixed input. Deterministic alternation, as used by the ring implementation, instead satisfies the two-round bound (154). The factor q/d is the cost of activating only a random coordinate subspace.

A.8 Proof of Theorem 3

Let $z_k = \Pi Y_t e_k$ be the centered network state on coordinate k. Expanding (42), then averaging over the matching and its supports, gives

$$\mathbb{E}\|\Pi X_{t+1}\|_F^2 = \|\Pi Y_t\|_F^2 - \frac{1}{2}\sum_{k=1}^{d} z_k^\top \bar{L} z_k. \tag{45}$$

Each $\bar{L}_k$ is a graph Laplacian. Hence

$$\sum_{k=1}^{d} z_k^\top \bar{L}_k z_k \geq \left(\min_{k \in [d]} \lambda_2(\bar{L}_k) \right) \sum_{k=1}^{d} \|z_k\|^2. \tag{46}$$

The coefficient is sharp: choose one nonzero column of Y_t to be a unit eigenvector associated with $\lambda_2(\bar{L}_k)$ for a minimizing coordinate k. This proves the equality in (15).

It remains to bound this coefficient by the scalar payload. Since $\mathrm{tr}(L_e) = 2$,

$$\sum_{k=1}^{d} \mathrm{tr}(\bar{L}_k) = 2\mathbb{E}\left[\sum_{e \in M_t} |S_{e,t}| \right] = C_{\mathrm{dir}}. \tag{47}$$

For any N-node Laplacian L, $\lambda_2(L) \leq \mathrm{tr}(L)/(N-1)$. Therefore

$$2\gamma_\star = \min_k \lambda_2(\bar{L}_k) \leq \frac{1}{d} \sum_{k=1}^{d} \lambda_2(\bar{L}_k) \leq \frac{C_{\mathrm{dir}}}{d(N-1)}, \tag{48}$$

which proves the converse and, after multiplying by B_{val}, the bit lower bound.

For a uniformly random perfect matching of K_N, every edge is selected with probability $1/(N-1)$, so

$$\mathbb{E}\left[\sum_{e \in M_t} L_e \right] = \frac{L_{K_N}}{N-1} = \frac{N}{N-1} \Pi. \tag{49}$$

A uniform rank-q support contains coordinate k with probability q/d, and a perfect matching sends $C_{\mathrm{dir}} = Nq$ directed values. Thus $\bar{L}_k = qN\Pi/[d(N-1)]$ for every k, and

$$\gamma_\star = \frac{qN}{2d(N-1)} = \frac{Nq}{2d(N-1)} = \frac{C_{\mathrm{dir}}}{2d(N-1)}, \tag{50}$$

To characterize equality, put $c = C_{\mathrm{dir}}/[d(N-1)]$. Equality in (48) requires all coordinate gaps to be equal, and all $N-1$ eigenvalues on the disagreement subspace of each $\bar{L}_k$ to equal its smallest one. Since $\bar{L}_k \mathbf{1} = 0$, these conditions are equivalent to

$$\bar{L}_k = c\Pi \quad \text{for every } k \in [d]. \tag{51}$$

Conversely, (51) gives $\gamma_\star = c/2$, attaining the bound. Unequal coordinate coverage or an anisotropic expected Laplacian makes at least one inequality strict. Uniform perfect matchings with uniform marginal supports satisfy this pure-mixing equality criterion for either common or independent edge supports.

A.9 Balanced Public Schedule

Let $K = \lceil d/q \rceil$. At epoch e, all agents use the shared seed to generate a public permutation π_e and take

$$\mathcal{S}_{eK+r} = \{\pi_e(rq+1), \ldots, \pi_e(\min\{(r+1)q, d\})\}, \qquad r = 0, \ldots, K-1. \tag{52}$$

The permutation π_e is a bijection from $[d]$ to $[d]$. The blocks in (52) form a partition of the ordered list

$$\pi_e(1), \pi_e(2), \ldots, \pi_e(d). \tag{53}$$

Hence every coordinate appears in at least one block of epoch e. If q divides d, every block has size exactly q, the blocks are disjoint, and every coordinate appears exactly once.

Since every coordinate appears at least once in every epoch, a coordinate activated in round $eK + r$ must be activated again by the end of the next epoch. The largest possible gap happens when the coordinate appears in the first round of epoch e and in the last round of epoch $e+1$, giving a gap of at most

$$(K-1) + K = 2K - 1. \tag{54}$$

The average refresh time is at most the epoch length K because each epoch contains at least one activation of the coordinate. The padding rule, when used, only adds extra activations and cannot increase the maximum gap or the average refresh time.

A.10 Masked Matrix Form of ZO-COSMO

Let $\mathbf{X}_t, \mathbf{G}_t \in \mathbb{R}^{N\times d}$ stack the row vectors $x_{i,t}$ and $\hat{g}_{i,t}$. Since every estimator is supported on $\mathcal{S}_t$, $\mathbf{G}_t P_t = \mathbf{G}_t$. The local update is

$$\mathbf{Y}_t = \mathbf{X}_t - \eta \mathbf{G}_t, \tag{55}$$

and the partial consensus step is

$$\mathbf{X}_{t+1} = \mathbf{Y}_t(I - P_t) + W\mathbf{Y}_t P_t. \tag{56}$$

Multiplying (56) by $N^{-1}\mathbf{1}^\top$ from the left and using $\mathbf{1}^\top W = \mathbf{1}^\top$ gives the exact averaged update

$$\bar{x}_{t+1} = \bar{x}_t - \eta \bar{g}_t, \qquad \bar{g}_t := \frac{1}{N}\sum_{i=1}^{N} \hat{g}_{i,t}. \tag{57}$$

Thus the exact descent direction of the network average is

$$h_t := \bar{g}_t. \tag{58}$$

A.11 Sparse Direction Identities

Lemma 2. *Let $\mathcal{S}$ be sampled uniformly from all q-subsets of $[d]$, let ϵ_j be independent Rademacher signs on $\mathcal{S}$, and define $u_j = \epsilon_j \mathbf{1}\{j \in \mathcal{S}\}$. Then*

$$\mathbb{E}[uu^\top] = \frac{q}{d} I_d, \qquad \|u\|^2 = q \quad \textit{almost surely}. \tag{59}$$

Consequently, for every deterministic vector $v \in \mathbb{R}^d$,

$$\mathbb{E}\left\| \frac{d}{q} uu^\top v \right\|^2 = d\|v\|^2. \tag{60}$$

Proof. For $j \neq k$, $\mathbb{E}[u_j u_k] = 0$ because the signs are independent and mean-zero conditional on the support. For $j = k$,

$$\mathbb{E}[u_j^2] = \mathbb{P}(j \in \mathcal{S}) = \frac{\binom{d-1}{q-1}}{\binom{d}{q}} = \frac{q}{d}. \tag{61}$$

This proves $\mathbb{E}[uu^\top] = (q/d) I_d$. Since exactly q entries of u are nonzero and each nonzero entry has magnitude one, $\|u\|^2 = q$. Therefore,

$$\begin{aligned} \mathbb{E}\left\| \frac{d}{q} uu^\top v \right\|^2 &= \left(\frac{d}{q}\right)^2 \mathbb{E}\big[(u^\top v)^2 \|u\|^2\big] \\ &= \left(\frac{d}{q}\right)^2 q\, v^\top \mathbb{E}[uu^\top] v = d\|v\|^2. \end{aligned} \tag{62}$$

□

Lemma 3. *Under the same sampling rule as Lemma 2, for any deterministic $v \in \mathbb{R}^d$,*

$$\mathbb{E}\big[(u^\top v)^4\big] \le 3\frac{q}{d}\|v\|^4, \tag{63}$$

Moreover,

$$\mathbb{E}\left\| \frac{d}{q} uu^\top v \right\|^4 \le 3\frac{d^3}{q}\|v\|^4. \tag{64}$$

Proof. Condition on the support $\mathcal{S}$. The random variable $u^\top v = \sum_{j\in\mathcal{S}} \varepsilon_j v_j$ is a Rademacher sum, so by the fourth-moment identity for independent signs,

$$\mathbb{E}_\varepsilon\big[(u^\top v)^4 \mid \mathcal{S}\big] = 3\left(\sum_{j\in\mathcal{S}} v_j^2\right)^2 - 2\sum_{j\in\mathcal{S}} v_j^4 \le 3\left(\sum_{j\in\mathcal{S}} v_j^2\right)^2. \tag{65}$$

For a uniform q-subset, the inclusion probabilities are $\mathbb{P}(j \in \mathcal{S}) = q/d$ and $\mathbb{P}(j,k \in \mathcal{S}) = q(q-1)/(d(d-1))$ for $j \neq k$. Therefore

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\left(\sum_{j\in\mathcal{S}} v_j^2\right)^2\right] &= \frac{q}{d}\sum_{j=1}^{d} v_j^4 + \frac{2q(q-1)}{d(d-1)}\sum_{1\le j<k\le d} v_j^2 v_k^2 \\ &\le \frac{q}{d}\left(\sum_{j=1}^{d} v_j^4 + 2\sum_{1\le j<k\le d} v_j^2 v_k^2\right) \\ &= \frac{q}{d}\|v\|^4.\end{aligned} \tag{66}$$

Combining (65) and (66) gives

$$\mathbb{E}\big[(u^\top v)^4\big] \le 3\frac{q}{d}\|v\|^4. \tag{67}$$

This proves (63). Since $\|u\|^2 = q$ almost surely,

$$\left\|\frac{d}{q}uu^\top v\right\|^4 = \left(\frac{d}{q}\right)^4 (u^\top v)^4\|u\|^4 = \left(\frac{d}{q}\right)^4 q^2 (u^\top v)^4. \tag{68}$$

Combining (68) with (63) gives

$$\mathbb{E}\left\|\frac{d}{q}uu^\top v\right\|^4 \le \left(\frac{d}{q}\right)^4 q^2 \cdot 3\frac{q}{d}\|v\|^4 = 3\frac{d^3}{q}\|v\|^4, \tag{69}$$

which is (64). □

Proposition 3. *Under Assumptions 3 and 4,*

$$\mathbb{E}[\hat{g}_{i,t} \mid x_{i,t}] = \nabla F_i(x_{i,t}) + b_\mu(x_{i,t}), \quad \|b_\mu(x_{i,t})\| \le B_\mu, \tag{70}$$

$$\mathbb{E}\|\hat{g}_{i,t}\|^2 \le M_{\text{node}}^2. \tag{71}$$

A.12 Proof of Proposition 3

Fix a node i, an iterate x, and write $u = u_t$. By Assumption 4,

$$\frac{f_i(x+\mu u;\xi) - f_i(x-\mu u;\xi)}{2\mu} = \nabla f_i(x;\xi)^\top u + r_\mu(x,u,\xi), \tag{72}$$

where

$$|r_\mu(x,u,\xi)| \le \frac{L_3}{6}\mu^2\|u\|^3. \tag{73}$$

Substituting (72) into (5) yields

$$\hat{g}_{i,t} = \underbrace{\frac{d}{q}uu^\top \nabla f_i(x;\xi)}_{a(x,u,\xi)} + \underbrace{\frac{d}{q}r_\mu(x,u,\xi)u}_{e_\mu(x,u,\xi)}. \tag{74}$$

By Lemma 2 and $\mathbb{E}_\xi[\nabla f_i(x;\xi)] = \nabla F_i(x)$,

$$\mathbb{E}[a(x,u,\xi) \mid x] = \frac{d}{q}\mathbb{E}[uu^\top]\nabla F_i(x) = \nabla F_i(x). \tag{75}$$

For the remainder, $\|u\|^2 = q$, so (73) gives

$$\|e_\mu(x,u,\xi)\| \le \frac{d}{q}\cdot\frac{L_3}{6}\mu^2\|u\|^3\cdot\|u\| = \frac{L_3}{6}\mu^2 dq = B_\mu. \tag{76}$$

Combining (75) and (76) proves (70).

For the second moment, first condition on ξ and apply Lemma 2:

$$\begin{aligned}\mathbb{E}_u\|a(x,u,\xi)\|^2 &= \left(\frac{d}{q}\right)^2 \mathbb{E}_u\left[(\nabla f_i(x;\xi)^\top u)^2\|u\|^2\right] \\ &= \left(\frac{d}{q}\right)^2 q\,\nabla f_i(x;\xi)^\top \mathbb{E}[uu^\top]\nabla f_i(x;\xi) \\ &= d\|\nabla f_i(x;\xi)\|^2.\end{aligned} \tag{77}$$

Taking expectation over ξ and using Assumption 3,

$$\mathbb{E}\|a(x,u,\xi)\|^2 \le d(G^2+\sigma^2). \tag{78}$$

Finally, $\|\hat{g}_{i,t}\|^2 \le 2\|a\|^2 + 2\|e_\mu\|^2$, so

$$\mathbb{E}\|\hat{g}_{i,t}\|^2 \le 2d(G^2+\sigma^2) + 2B_\mu^2 = M_{\text{node}}^2. \tag{79}$$

This proves (71).

A.13 Averaged Estimator Bounds

Lemma 4. *Define*

$$\bar{g}_t := \frac{1}{N}\sum_{i=1}^N \hat{g}_{i,t}, \qquad \bar{\nabla}_t := \frac{1}{N}\sum_{i=1}^N \nabla F_i(x_{i,t}). \tag{80}$$

Let $\mathcal{F}_t$ denote the sigma-field generated by all iterates and all randomness before the support, direction, and oracle samples at time t are drawn. Under Assumptions 1–4,

$$\mathbb{E}\|\bar{g}_t\|^2 \le Cd\,\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + CdL^2\mathcal{E}_t + C\frac{d\sigma^2}{N} + CB_\mu^2, \tag{81}$$

More precisely, before replacing the local average gradient by the gradient at the network average,

$$\mathbb{E}\big[\|\bar{g}_t\|^2 \mid \mathcal{F}_t\big] \le Cd\,\|\bar{\nabla}_t\|^2 + C\frac{d\sigma^2}{N} + CB_\mu^2. \tag{82}$$

Moreover,

$$\mathbb{E}\big[\langle \nabla F(\bar{x}_t), \bar{g}_t\rangle\big] \ge \frac{3}{4}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 - CL^2\mathcal{E}_t - CB_\mu^2. \tag{83}$$

The only variance term reduced by N is the independent oracle noise. The sparse direction u_t is public across agents, so its random-direction variance is shared rather than averaged away.

Proof. By Jensen's inequality and L-smoothness,

$$\begin{aligned}\mathbb{E}\|\bar{\nabla}_t - \nabla F(\bar{x}_t)\|^2 &= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N\big(\nabla F_i(x_{i,t}) - \nabla F_i(\bar{x}_t)\big)\right\|^2 \\ &\le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\nabla F_i(x_{i,t}) - \nabla F_i(\bar{x}_t)\|^2 \\ &\le L^2\mathcal{E}_t.\end{aligned} \tag{84}$$

The price of decentralization is visible here: local gradients are evaluated at $x_{i,t}$ rather than at the network average $\bar{x}_t$.

Next write the stochastic gradient noise as

$$\zeta_{i,t} := \nabla f_i(x_{i,t};\xi_{i,t}) - \nabla F_i(x_{i,t}), \qquad \mathbb{E}[\zeta_{i,t} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}\|\zeta_{i,t}\|^2 \le \sigma^2. \tag{85}$$

The samples $\xi_{i,t}$ are independent across agents, hence

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \zeta_{i,t}\right\|^2 \middle| \mathcal{F}_t\right] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\|\zeta_{i,t}\|^2 \le \frac{\sigma^2}{N}. \tag{86}$$

Equation (86) reduces independent oracle variance by N. The common direction u_t remains in the averaged estimator.

Using (74), the average estimator can be written as

$$\bar{g}_t = \frac{d}{q} u_t u_t^\top \left(\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{i,t}; \xi_{i,t}) \right) + \frac{1}{N} \sum_{i=1}^{N} e_{\mu,i,t}. \tag{87}$$

By Lemma 2 and the conditional variance identity (86),

$$\mathbb{E}\big[\|\bar{g}_t\|^2 \mid \mathcal{F}_t\big] \le Cd\,\|\bar{\nabla}_t\|^2 + C\frac{d\sigma^2}{N} + CB_\mu^2, \tag{88}$$

which is (82). Taking expectation and using (84) gives

$$\begin{aligned} \mathbb{E}\|\bar{g}_t\|^2 &\le Cd\,\mathbb{E}\|\bar{\nabla}_t\|^2 + C\frac{d\sigma^2}{N} + CB_\mu^2 \\ &\le Cd\,\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + CdL^2\mathcal{E}_t + C\frac{d\sigma^2}{N} + CB_\mu^2. \end{aligned} \tag{89}$$

This proves (81).

We also need an alignment bound. Conditional on $\mathcal{F}_t$, Proposition 3 gives

$$\mathbb{E}[\bar{g}_t \mid \mathcal{F}_t] = \bar{\nabla}_t + \bar{b}_{\mu,t}, \qquad \|\bar{b}_{\mu,t}\| \le B_\mu. \tag{90}$$

Using Young's inequality twice,

$$\begin{aligned} \mathbb{E}\big[\langle \nabla F(\bar{x}_t), \bar{g}_t \rangle\big] &= \mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + \mathbb{E}\langle \nabla F(\bar{x}_t), \bar{\nabla}_t - \nabla F(\bar{x}_t) \rangle \\ &\quad + \mathbb{E}\langle \nabla F(\bar{x}_t), \bar{b}_{\mu,t} \rangle \\ &\ge \frac{3}{4}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 - CL^2\mathcal{E}_t - CB_\mu^2. \end{aligned} \tag{91}$$

This proves (83). □

A.14 Base Direction for the Convergence Proof

By (57), the direction used by Algorithm 1 is exactly

$$h_t = \bar{g}_t. \tag{92}$$

The second-moment and alignment bounds for the actual descent direction are therefore

$$\mathbb{E}\|h_t\|^2 \le Cd\,\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + CdL^2\mathcal{E}_t + C\frac{d\sigma^2}{N} + CB_\mu^2, \tag{93}$$

$$\mathbb{E}\langle \nabla F(\bar{x}_t), h_t \rangle \ge \frac{3}{4}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 - CL^2\mathcal{E}_t - CB_\mu^2. \tag{94}$$

Both displays follow immediately from Lemma 4 and (92).

A.15 Masked Disagreement Energy

Lemma 5. *For Algorithm 1, the disagreement sequence satisfies the one-step recursion* (102) *and the averaged bound* (104)*.*

Proof. Let $J = I_N - N^{-1}\mathbf{1}\mathbf{1}^\top$. Then $\mathcal{E}_t = N^{-1}\mathbb{E}\|J\mathbf{X}_t\|_F^2$. Consider coordinate j. If $j \notin \mathcal{S}_t$, then $y^{(j)} = x^{(j)}$ and the coordinate is not mixed. If $j \in \mathcal{S}_t$, then $y^{(j)} = x^{(j)} - \eta g^{(j)}$, and the update is mixed by W. Hence

$$\|JWy^{(j)}\|^2 \le \rho^2 \|Jy^{(j)}\|^2, \tag{95}$$

where $x^{(j)}, g^{(j)}, y^{(j)}$ are the j-th columns of $\mathbf{X}_t, \mathbf{G}_t, \mathbf{Y}_t$. On active coordinates, the relaxed triangle inequality gives

$$\|Jy^{(j)}\|^2 \le (1+\alpha)\|Jx^{(j)}\|^2 + (1+\alpha^{-1})\eta^2\|Jg^{(j)}\|^2. \tag{96}$$

Condition on $\mathcal{F}_t$ and average over the fresh public coin and oracle samples. Since $g^{(j)} = 0$ off the active support and $\mathbb{P}(j \in \mathcal{S}_t) = q/d$,

$$\begin{aligned} \mathbb{E}\big[\|Jx_{t+1}^{(j)}\|^2 \mid \mathcal{F}_t\big] &\le \Big(1 - \frac{q}{d}\Big)\|Jx^{(j)}\|^2 + \frac{q}{d}\rho^2(1+\alpha)\|Jx^{(j)}\|^2 \\ &\quad + \rho^2(1+\alpha^{-1})\eta^2\mathbb{E}\big[\|Jg^{(j)}\|^2 \mid \mathcal{F}_t\big]. \end{aligned} \tag{97}$$

If $\rho = 0$, the active-coordinate mixing term vanishes and the same bound follows by taking the limit below. Assume $\rho > 0$ and choose $\alpha = (1-\rho^2)/(2\rho^2)$. Then $\rho^2(1+\alpha) = (1+\rho^2)/2$, and therefore

$$\left(1 - \frac{q}{d}\right) + \frac{q}{d}\rho^2(1+\alpha) = 1 - \frac{q}{d}\frac{1-\rho^2}{2}. \tag{98}$$

With the same choice of α,

$$\rho^2(1+\alpha^{-1}) = \rho^2 + \frac{2\rho^4}{1-\rho^2} \le \frac{C}{1-\rho^2}. \tag{99}$$

Summing (97) over coordinates and taking the remaining expectation yields

$$\mathcal{E}_{t+1} \le \left(1 - \frac{q}{d}\frac{1-\rho^2}{2}\right)\mathcal{E}_t + \frac{C\eta^2}{1-\rho^2}\mathbb{E}\left[\frac{1}{N}\|J\mathbf{G}_t\|_F^2\right]. \tag{100}$$

By $\|J\|_2 \le 1$ and Proposition 3,

$$\mathbb{E}\left[\frac{1}{N}\|J\mathbf{G}_t\|_F^2\right] \le \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\hat{g}_{i,t}\|^2 \le M_{\text{node}}^2. \tag{101}$$

Therefore,

$$\mathcal{E}_{t+1} \le \kappa\mathcal{E}_t + \frac{C\eta^2}{1-\rho^2}M_{\text{node}}^2, \qquad \kappa := 1 - \frac{q}{d}\frac{1-\rho^2}{2}. \tag{102}$$

Unrolling (102) gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t \le \frac{\mathcal{E}_0}{T(1-\kappa)} + \frac{C\eta^2 M_{\text{node}}^2}{(1-\rho^2)(1-\kappa)}. \tag{103}$$

Since $1 - \kappa = (q/d)(1-\rho^2)/2$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t \le C\frac{d}{q}\frac{\mathcal{E}_0}{T(1-\rho^2)} + C\eta^2\frac{d}{q}\frac{M_{\text{node}}^2}{(1-\rho^2)^2}. \tag{104}$$

□

The second term in (104) is the sparse-consensus term used in the main theorem. The first term vanishes when all agents are initialized equally and is otherwise a transient.

Sparse-consensus factor. Equation (104) is where the random subspace design enters the network analysis. Full-coordinate mixing contracts disagreement energy at rate $1-\rho^2$. ZO-COSMO mixes coordinate j only on the event $j \in S_t$, which has probability q/d. The effective energy contraction probability is therefore of order $(q/d)(1-\rho^2)$, and unrolling the recursion produces the multiplier d/q.

The second spectral-gap factor is already visible in (99). The active-coordinate descent step perturbs consensus before mixing, and bounding that perturbation costs one factor of $(1-\rho^2)^{-1}$. Unrolling the masked contraction costs another. Together they give the squared energy gap in Theorem 4, along with the sparse activation factor d/q.

The q/d factor is intrinsic to this one-step partial-consensus model, as the next lemma shows. The remaining slack enters through Young's inequality in (99) and the bound of the injected disagreement by the full per-node estimator moment. For matching updates, Proposition 1 instead separates aggregate estimation error from disagreement injection, exposing the effect of direction correlation.

Lemma 6. *The sparse activation factor q/d in* (102) *is tight for partial coordinate mixing. In particular, there are disagreement states for which one pure partial-consensus step with $\eta = 0$ contracts the expected disagreement energy by exactly*

$$1 - \frac{q}{d}(1-\rho^2). \tag{105}$$

Proof. Because W is symmetric and doubly stochastic, there is a unit vector $v \perp \mathbf{1}$ and an eigenvalue λ with $Wv = \lambda v$ and $|\lambda| = \rho$. Take a disagreement matrix whose j-th column is $x^{(j)} = a_j v$ for

arbitrary coefficients a_j, and set $\eta = 0$ so that the step only applies partial consensus. If $j \notin S_t$, then $x_+^{(j)} = x^{(j)}$. If $j \in S_t$, then $x_+^{(j)} = Wx^{(j)} = a_j \lambda v$, and hence

$$\|Jx_+^{(j)}\|^2 = \begin{cases} \|x^{(j)}\|^2, & j \notin S_t, \\ \rho^2 \|x^{(j)}\|^2, & j \in S_t. \end{cases} \tag{106}$$

Since $\mathbb{P}(j \in S_t) = q/d$,

$$\mathbb{E}\|Jx_+^{(j)}\|^2 = \left(1 - \frac{q}{d}(1-\rho^2)\right) \|x^{(j)}\|^2. \tag{107}$$

Summing over coordinates gives (105). Thus, for algorithms that mix only the active coordinates and leave inactive coordinates unchanged, the q/d slowdown in the one-step consensus contraction cannot be removed by sharper analysis alone. □

A.16 Proof of Theorem 4

By Assumption 1 and the averaged update (57),

$$F(\bar{x}_{t+1}) \le F(\bar{x}_t) - \eta \langle \nabla F(\bar{x}_t), h_t \rangle + \frac{L\eta^2}{2} \|h_t\|^2. \tag{108}$$

Taking expectations and substituting (94) and (93) gives

$$\begin{aligned} \mathbb{E}F(\bar{x}_{t+1}) \le \mathbb{E}F(\bar{x}_t) - \frac{3\eta}{4}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + C\eta L^2 \mathcal{E}_t + C\eta B_\mu^2 \\ + CL\eta^2 d\, \mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + CL\eta^2 d L^2 \mathcal{E}_t \\ + CL\eta^2 \frac{d\sigma^2}{N} + CL\eta^2 B_\mu^2. \end{aligned} \tag{109}$$

Choose the universal constant in (17) so that $CL\eta d \le 1/4$. The positive stationarity term is then at most $(\eta/4)\mathbb{E}\|\nabla F(\bar{x}_t)\|^2$. The same condition absorbs $CL\eta d$ into the disagreement coefficient, and $\eta \le L^{-1}$ absorbs the smaller bias term. Thus

$$\mathbb{E}F(\bar{x}_{t+1}) \le \mathbb{E}F(\bar{x}_t) - \frac{\eta}{2}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 + C\eta L^2 \mathcal{E}_t + C\eta B_\mu^2 + CL\eta^2 \frac{d\sigma^2}{N}. \tag{110}$$

Summing (110) from $t = 0$ to $T-1$ and using $F(\bar{x}_T) \ge F^*$ gives

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \le \frac{2(F(\bar{x}_0) - F^*)}{\eta T} + CL^2 \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{E}_t \\ + CB_\mu^2 + CL\eta \frac{d\sigma^2}{N}. \end{aligned} \tag{111}$$

Substitute the masked disagreement bound from Lemma 5. This gives

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \le \frac{C(F(\bar{x}_0) - F^*)}{\eta T} + CL\eta \frac{d\sigma^2}{N} + CB_\mu^2 \\ + CL^2\eta^2 \frac{d}{q} \frac{M_{\text{node}}^2}{(1-\rho^2)^2} + CL^2 \frac{d}{q} \frac{\mathcal{E}_0}{T(1-\rho^2)}, \end{aligned} \tag{112}$$

Using $B_\mu^2 = \mathcal{O}(\mu^4 d^2 q^2)$ and $M_{\text{node}}^2 = \mathcal{O}(d(G^2 + \sigma^2) + \mu^4 d^2 q^2)$ gives the explicit finite-η form

$$\begin{aligned} R_T &:= \frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \\ &\le \mathcal{O}\Bigg(\frac{\Delta_0}{\eta T} + \frac{\eta d \sigma^2}{N} + \mu^4 d^2 q^2 \\ &\qquad + \frac{\eta^2 d[d(G^2 + \sigma^2) + \mu^4 d^2 q^2]}{q(1-\rho^2)^2} + \frac{d\mathcal{E}_0}{qT(1-\rho^2)} \Bigg). \end{aligned} \tag{113}$$

With common initialization and (18), for any feasible η,

$$R_T \le \mathcal{O}\left(\frac{\Delta_0}{\eta T} + \eta \frac{d\sigma^2}{N} + \frac{1}{T} + \eta^2 \frac{d^2(G^2 + \sigma^2)}{q(1-\rho^2)^2} \right). \tag{114}$$

Taking $\eta = c_\eta T^{-1/2}$ in (114), for horizons satisfying (17), gives (19). The directed-link form (214) then follows from the round substitution in Appendix C.7. If $\sigma^2 > 0$ and

$$\eta_\star = \Theta\left(\sqrt{\frac{N\Delta_0}{d\sigma^2 T}}\right) \tag{115}$$

satisfies (17), then

$$R_T \le \mathcal{O}\left(\sqrt{\frac{\Delta_0 d\sigma^2}{NT}} + \frac{1}{T} + \frac{N\Delta_0 d(G^2+\sigma^2)}{q(1-\rho^2)^2\sigma^2 T}\right). \tag{116}$$

A.17 Descent Absorption

We spell out the passage from (109) to (110). The only positive term containing the stationarity measure is

$$CL\eta^2 d\,\mathbb{E}\|\nabla F(\bar{x}_t)\|^2. \tag{117}$$

Choose the universal constant c in (17) so that $CL\eta d \le 1/4$. Then (117) is at most $(\eta/4)\mathbb{E}\|\nabla F(\bar{x}_t)\|^2$. Combining it with the coefficient $-3\eta/4$ in (109) leaves

$$-\frac{\eta}{2}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2. \tag{118}$$

The two disagreement terms factor as

$$C\eta L^2\mathcal{E}_t + CL\eta^2 dL^2\mathcal{E}_t = C\eta L^2\mathcal{E}_t(1+L\eta d), \tag{119}$$

which is at most $C'\eta L^2\mathcal{E}_t$ under the same condition. Likewise, since $L\eta \le 1$,

$$C\eta B_\mu^2 + CL\eta^2 B_\mu^2 \le C'\eta B_\mu^2. \tag{120}$$

The remaining stochastic term is $CL\eta^2 d\sigma^2/N$. After summing over t and dividing by ηT, it becomes $CL\eta d\sigma^2/N$. These are precisely the three positive terms in (110); all graph dependence enters later through (104).

A.18 Worker-Scaled Rate Derivation

For common initialization, $\mathcal{E}_0 = 0$. Equation (116) follows by setting

$$A_q(\mu) := \frac{d}{q(1-\rho^2)^2}\left[d(G^2+\sigma^2) + \mu^4 d^2 q^2\right]. \tag{121}$$

Ignoring only constants depending on L, (113) can be written as

$$R_T(\eta,\mu) := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \le \mathcal{O}\left(\frac{\Delta_0}{\eta T}\right) + \mathcal{O}\left(\eta\frac{d\sigma^2}{N}\right) + \mathcal{O}\big(\mu^4 d^2 q^2\big) + \mathcal{O}\big(\eta^2 A_q(\mu)\big). \tag{122}$$

The first term is the usual descent term, the second term is the averaged stochastic ZO noise, the third term is the finite-difference bias, and the fourth term is the sparse decentralized consensus penalty.

Balancing the first two terms in (122) gives

$$\frac{\Delta_0}{\eta T} \asymp \eta\frac{d\sigma^2}{N}, \qquad \eta_\star \asymp \sqrt{\frac{N\Delta_0}{d\sigma^2 T}}, \tag{123}$$

Substituting (123) back into the first two terms yields

$$\frac{\Delta_0}{\eta_\star T} + \eta_\star\frac{d\sigma^2}{N} = \mathcal{O}\left(\sqrt{\frac{\Delta_0 d\sigma^2}{NT}}\right). \tag{124}$$

The sparse-consensus term becomes

$$\eta_\star^2 A_q(\mu) = \mathcal{O}\left(\frac{N\Delta_0}{d\sigma^2 T}\frac{d}{q(1-\rho^2)^2}\left[d(G^2+\sigma^2)+\mu^4 d^2 q^2\right]\right). \tag{125}$$

Using the canonical smoothing choice

$$\mu = \Theta\Big((dq)^{-1/2}T^{-1/4}\Big) \tag{126}$$

gives

$$\mu^4 d^2 q^2 = \mathcal{O}(T^{-1}) \tag{127}$$

and therefore the finite-difference bias contributes $\mathcal{O}(T^{-1})$. The bias-dependent part inside (125) is also of smaller order than the displayed consensus term for the regimes used in our experiments. Thus

$$\eta_\star^2 A_q(\mu) = \mathcal{O}\left(\frac{N\Delta_0 d(G^2+\sigma^2)}{q(1-\rho^2)^2\sigma^2 T}\right) + \text{higher-order smoothing terms.} \tag{128}$$

Combining (124), (127), and (128) gives

$$R_T \le \mathcal{O}\left(\sqrt{\frac{\Delta_0 d\sigma^2}{NT}}\right) + \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{N\Delta_0 d(G^2+\sigma^2)}{q(1-\rho^2)^2\sigma^2 T}\right). \tag{129}$$

The same public coordinates are used across agents, so worker averaging removes independent oracle noise but not sparse-coordinate sampling itself.

A.19 Parameter Choice

The smoothing radius μ controls finite-difference bias, the stepsize η balances descent against stochastic and network error, and the sparsity q controls communication and coordinate refresh.

For a central-difference estimator with a third-order smoothness bound, the bias scales as

$$B_\mu^2 = \mathcal{O}(\mu^4 d^2 q^2). \tag{130}$$

The dependence on q comes from $\|u_t\|^4 = q^2$ and the scaling factor d/q used to keep the sparse estimator unbiased to first order. If q is increased while μ is held fixed, the finite-difference truncation error can grow. The canonical theoretical choice therefore sets

$$\mu = \Theta\left(\frac{1}{\sqrt{dq}\,T^{1/4}}\right), \tag{131}$$

which keeps (130) at order T^{-1}. The experimental values of μ are fixed separately to accommodate loss scale, minibatch noise, and numerical precision. The theoretical scaling prescribes a smaller radius as q or d increases.

The stepsize has two competing constraints. The smooth descent step requires

$$\eta \le c(Ld)^{-1} \tag{132}$$

for a universal constant c. The stochastic rate suggests the larger balancing choice

$$\eta \asymp \sqrt{\frac{N\Delta_0}{d\sigma^2 T}}, \tag{133}$$

whenever this value satisfies (132). In high-dimensional full-model tuning, the stability constraint can be much more restrictive than the formal stochastic balancing rule. This is consistent with the Qwen full-model study: the estimator is multiplied by d/q, so a learning rate that is harmless in a low-dimensional LoRA subspace can be unstable when the trainable coordinate set has billions of dimensions.

The sparsity level q creates the main design tradeoff. Per round, larger q improves coordinate coverage and reduces the sparse-consensus factor d/q, but it also increases the value payload linearly. Under a fixed directed-link budget $\mathcal{B}$, the number of affordable rounds is approximately

$$T(q) \approx \frac{\mathcal{B}}{qB_{\text{val}}}. \tag{134}$$

Substitution into (214) cancels the explicit $1/q$ factor in the network term. The remaining displayed terms increase with q when the other constants and the prescribed horizon-dependent parameters are held fixed. This bound therefore favors smaller supports on communication alone; it does not select an interior optimum. Practical selection also accounts for the two forward queries per round, their latency, and finite-difference precision.

For iid uniform supports at a node, coordinate exposure is measured by $\tau = qT/d$. For any fixed coordinate j,

$$\mathbb{E}\sum_{t<T}\mathbf{1}\{j \in S_t\} = \frac{qT}{d}, \qquad \Pr\{j \notin \cup_{t<T} S_t\} = (1-q/d)^T. \tag{135}$$

The first identity sums the activation probabilities; the second uses independence across rounds. Hence d/q is the mean waiting time, not a deterministic coverage period. At a fixed value-only budget, $\tau \approx \mathcal{B}/(dB_{\mathrm{val}})$, independent of q to first order. The support-size sweep in Table 9 compares the resulting optimization trajectories at matched bits.

The graph parameter ρ enters through the disagreement recursion, producing $(1-\rho^2)^{-2}$ in the bound. The one-step witness in Appendix A.15 establishes the q/d contraction factor; it does not establish sharpness of the full spectral dependence. Sparse momentum retains information between coordinate activations. Theorem 5 gives its convergence guarantee, and Appendix B.10 tests the interaction between its step size and query correlation.

A.20 Convergence with Sparse Momentum

The finite difference and local update use the current mixing support. We apply contraction to the pre-query state and bound the resulting update separately. This argument covers global support, independent supports on matched edges, and a common support on all matched edges.

Write $\mathcal{A}_t$ for the linear mixing operator and $\Pi = I - N^{-1}\mathbf{1}\mathbf{1}^\top$. Each $\mathcal{A}_t$ preserves the average, commutes with Π, and is nonexpansive in Frobenius norm. Suppose that for an integer $h \geq 1$ and $\gamma \in (0,1]$, every block satisfies

$$\mathbb{E}\big[\|\mathcal{A}_{kh+h-1}\cdots\mathcal{A}_{kh}Z\|_F^2 \mid \mathcal{F}_{kh}\big] \leq (1-\gamma)\|Z\|_F^2 \qquad (\mathbf{1}^\top Z = 0), \tag{136}$$

for every $\mathcal{F}_{kh}$-measurable Z. The block's supports and matching schedule are independent of the state at its start. Their dependence on subsequent within-block iterates is handled explicitly below. Every node is active, and its rank-q projector $P_{i,t}$ is conditionally uniform given the past. Put $p = q/d$ and $a_\beta = \beta/(1-\beta)$.

Theorem 5. *Under Assumptions 1, 3, and 4, consider either global or perfect-matching ZO-COSMO with* (3)*, common initialization, and zero initial momentum. If* (136) *holds, then for any constant* $\eta > 0$*,* $\mu > 0$*, and* $0 \leq \beta < 1$*,*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \leq \mathcal{O}\left(\frac{\Delta_0}{\eta T} + L\eta M_{\mathrm{node}}^2 + B_\mu^2 + L^2\eta^2 M_{\mathrm{node}}^2\left[\frac{h^2}{\gamma^2} + \frac{a_\beta^2}{p}\right]\right). \tag{137}$$

The hidden constant is numerical. Choosing $\eta = \Theta(T^{-1/2})$ *and* $\mu = \Theta((dq)^{-1/2}T^{-1/4})$ *yields vanishing average stationarity for fixed problem parameters,* h, γ*, and* β*. A directed-link or per-node budget is incorporated by substituting the number of rounds affordable under that protocol's actual payload.*

Proof. Let G_t, M_t, V_t stack the gradient estimates, stored momentum, and applied directions. On node i, the update is

$$m_{i,t+1} = (I - P_{i,t})m_{i,t} + \beta P_{i,t}m_{i,t} + (1-\beta)\hat{g}_{i,t}, \qquad v_{i,t} = P_{i,t}m_{i,t+1}. \tag{138}$$

Since $P_{i,t}\hat{g}_{i,t} = \hat{g}_{i,t}$, rearranging gives the exact identity

$$v_{i,t} = \hat{g}_{i,t} - a_\beta(m_{i,t+1} - m_{i,t}). \tag{139}$$

This identity retains the inactive entries of $m_{i,t}$ and requires no independence between the query and the projector.

First control the moments. Orthogonality of active and inactive coordinates and convexity of the squared norm give

$$\mathbb{E}\|m_{i,t+1}\|^2 \leq [1-(1-\beta)p]\mathbb{E}\|m_{i,t}\|^2 + (1-\beta)M_{\mathrm{node}}^2, \tag{140}$$

$$\mathbb{E}\|v_{i,t}\|^2 \leq \beta p\mathbb{E}\|m_{i,t}\|^2 + (1-\beta)M_{\mathrm{node}}^2. \tag{141}$$

Here only the old momentum is conditioned on the past; $\mathbb{E}\|P_{i,t}m_{i,t}\|^2 = p\mathbb{E}\|m_{i,t}\|^2$. Proposition 3 bounds the fresh estimate, including its dependence on the current support. Starting from zero

momentum, induction proves

$$\mathbb{E}\|m_{i,t}\|^2 \le M_{\mathrm{node}}^2/p, \qquad \mathbb{E}\|v_{i,t}\|^2 \le M_{\mathrm{node}}^2, \qquad \mathbb{E}\|\bar{g}_t\|^2 \le M_{\mathrm{node}}^2. \tag{142}$$

The last bound follows from Jensen's inequality for both common and edge-local directions. It uses the per-node second moment, without a worker-averaging factor.

Next bound disagreement. The actual recursion is

$$X_{t+1} = \mathcal{A}_t(X_t - \eta V_t). \tag{143}$$

Over a block, expand it as

$$\Pi X_{kh+h} = \mathcal{A}_{kh+h-1}\cdots\mathcal{A}_{kh}\Pi X_{kh} - \eta\sum_{s=0}^{h-1}\mathcal{A}_{kh+h-1}\cdots\mathcal{A}_{kh+s}\Pi V_{kh+s}. \tag{144}$$

Nonexpansiveness and $\|\sum_{s=0}^{h-1} b_s\|^2 \le h\sum_s \|b_s\|^2$ bound the normalized expected squared norm of the second term by $\eta^2h^2M_{\mathrm{node}}^2$. This bound is pathwise before expectation and remains valid when V_{kh+s} depends on any of the block's supports. Apply (136) only to the first term. For $\gamma < 1$, Young's inequality with $\alpha = \gamma/[2(1-\gamma)]$ yields

$$\mathcal{E}_{kh+h} \le (1-\gamma/2)\mathcal{E}_{kh} + \frac{3\eta^2h^2M_{\mathrm{node}}^2}{\gamma}. \tag{145}$$

If $\gamma = 1$, the homogeneous term has zero second moment and the same inequality follows directly. Common initialization gives $\mathcal{E}_{kh} \le 6\eta^2h^2M_{\mathrm{node}}^2/\gamma^2$. Expanding a partial block of length $r < h$ and using the same norm bound gives, including the final incomplete block,

$$\sup_{t<T}\mathcal{E}_t \le C\eta^2h^2M_{\mathrm{node}}^2/\gamma^2. \tag{146}$$

Finally, define the virtual network average

$$z_t = \bar{x}_t - \eta a_\beta \bar{m}_t. \tag{147}$$

Average preservation and (139) imply $z_{t+1} = z_t - \eta\bar{g}_t$ exactly, with $z_0 = \bar{x}_0$. The conditional mean of $\bar{g}_t$ is $N^{-1}\sum_i \nabla F_i(x_{i,t}) + \bar{b}_t$, where $\|\bar{b}_t\| \le B_\mu$. Smoothness, Jensen's inequality, and (142) therefore imply

$$\mathbb{E}\,\|\mathbb{E}[\bar{g}_t \mid \mathcal{F}_t] - \nabla F(z_t)\|^2 \le CL^2\mathcal{E}_t + CL^2\eta^2a_\beta^2M_{\mathrm{node}}^2/p + CB_\mu^2. \tag{148}$$

Applying smooth descent at z_t, and using $\langle a,b\rangle \ge \|a\|^2/2 - \|a-b\|^2/2$, gives

$$\begin{aligned}\mathbb{E}F(z_{t+1}) \le \mathbb{E}F(z_t) - \frac{\eta}{2}\mathbb{E}\|\nabla F(z_t)\|^2 + C\eta L^2\mathcal{E}_t + C\eta B_\mu^2 \\ + CL^2\eta^3a_\beta^2M_{\mathrm{node}}^2/p + \frac{L\eta^2}{2}M_{\mathrm{node}}^2.\end{aligned} \tag{149}$$

Sum over $t < T$, use $F(z_T) \ge F^*$, and divide by $\eta T/2$. The inequality

$$\mathbb{E}\|\nabla F(\bar{x}_t)\|^2 \le 2\mathbb{E}\|\nabla F(z_t)\|^2 + 2L^2\eta^2a_\beta^2M_{\mathrm{node}}^2/p \tag{150}$$

then transfers the result back to the reported network average. Substitution of (146) proves (137). □

Schedules covered by the bound. For global support, the frozen-input calculation gives $h = 1$ and $\gamma = p(1-\rho^2)$. For independent perfect matchings with $\lambda_2(\bar{L}) > 0$, Theorem 2 gives $h = 1$ and $\gamma = p\lambda_2(\bar{L})/2$. A common uniform support on all edges of the matching gives the same pure-mixing expectation: independence across disjoint edge supports is not needed for that identity.

The implemented ring schedule for even $N \ge 4$ alternates two perfect matchings. Each phase is disconnected, so we bound contraction over two rounds. Write L_0, L_1 for the matching Laplacians; $L_0 + L_1 = L_{\mathrm{ring}}$. For a fixed centered Z, let $Z' = \mathcal{A}_0 Z$ and define the expected energy drops

$$D_0 = \frac{p}{2}\operatorname{tr}(Z^\top L_0 Z), \qquad D_1 = \frac{p}{2}\mathbb{E}\operatorname{tr}(Z'^\top L_1 Z'). \tag{151}$$

The second support is fresh, so the total two-round drop is $D_0 + D_1$. Moreover, $\mathbb{E}\|Z - Z'\|_F^2 = D_0$ because the first mixing step is an orthogonal projection. Since $\|L_1\|_2 = 2$,

$$\operatorname{tr}(Z^\top L_1 Z) \le 2\mathbb{E}\operatorname{tr}(Z'^\top L_1 Z') + 4D_0. \tag{152}$$

Consequently,

$$\frac{p}{2}\operatorname{tr}(Z^\top L_{\mathrm{ring}} Z) \le (1+2p)D_0 + 2D_1 \le 3(D_0+D_1). \tag{153}$$

Thus (136) holds with

$$h = 2, \qquad \gamma = \frac{p\lambda_2(L_{\mathrm{ring}})}{6} = \frac{p}{3}\big(1-\cos(2\pi/N)\big). \tag{154}$$

Substituting this block contraction into (137) establishes convergence for alternating-ring momentum updates with either common or independent edge supports. Direction-dependent moment bounds are developed in Appendix B.2.

B Correlation and Optimization

We analyze Algorithm 2 without momentum, holding the matching schedule fixed across couplings. Correlated compression already distinguishes individual from aggregate estimator error (Szlendak et al., 2022). Here the same direction also determines which peer-state coordinates are mixed, so aggregate variance must be analyzed together with the disagreement update. The identities below describe that joint operation.

In this section $\mathcal{D}_t = N^{-1}\|\Pi X_t\|_F^2$ is the instantaneous state disagreement; its expectation is the $\mathcal{E}_t$ used in the main convergence analysis.

B.1 Protocol and Couplings

Let $N = 2m$ nodes maintain $x_i \in \mathbb{R}^d$. A perfect matching M is sampled independently of the states. Each pair $e = \{i, j\}$ uses a uniform q-coordinate support S_e, signs ϵ_e, and

$$P_e = \operatorname{diag}(\mathbf{1}_{S_e}), \qquad u_e = P_e\epsilon_e, \qquad D_e = \frac{d}{q}u_e u_e^\top, \qquad p = \frac{q}{d}. \tag{155}$$

Both endpoints use the same u_e. Across distinct pairs we consider:

1. G: one common support and one common sign vector;
2. S: one common support, with independent sign vectors across pairs;
3. I: independent supports and sign vectors across pairs.

The public seed, round counter, and unordered edge identifiers generate all three choices without recurring support metadata. Each node evaluates

$$\hat{g}_i = \frac{d}{q}\frac{F_i(x_i+\mu u_e) - F_i(x_i-\mu u_e)}{2\mu}u_e, \qquad y_i = x_i - \eta\hat{g}_i, \tag{156}$$

and sends q ordered values of y_i to its matched peer. The endpoints average those coordinates and retain their own remaining entries. Every variant therefore uses two optimization queries and qB_{val} outgoing value bits per node per round. Initial seed setup is separate.

Each run keeps its coupling fixed; support and sign draws are fresh at every round.

B.2 Conditional Moments

Fix the states and matching before drawing directions. Put

$$a_e = \frac{\nabla F_i(x_i) + \nabla F_j(x_j)}{2}, \quad \bar{a} = \frac{1}{m}\sum_e a_e = \frac{1}{N}\sum_i \nabla F_i(x_i), \quad H_M^2 = \frac{1}{m}\sum_e \|a_e - \bar{a}\|^2. \tag{157}$$

For the ideal directional estimates let $v_e = D_e a_e$ and $\bar{v} = m^{-1}\sum_e v_e$. Define

$$V_s = \mathbb{E}\,\|\bar{v} - \bar{a}\|^2, \qquad U_s = \frac{1}{m}\sum_e \mathbb{E}\,\|v_e - \bar{v}\|^2, \qquad s \in \{\mathrm{G}, \mathrm{S}, \mathrm{I}\}. \tag{158}$$

Here U_s measures disagreement injected by the applied pair updates, not the current state disagreement.

For $1 \le q \le d$, the three couplings satisfy

$$V_s = \alpha_s \|\bar{a}\|^2 + \tau_s H_M^2, \tag{159}$$

$$U_s = (d - 1 - \alpha_s) \|\bar{a}\|^2 + (d - \tau_s) H_M^2, \tag{160}$$

where

$$\begin{aligned} (\alpha_{\mathrm{G}}, \tau_{\mathrm{G}}) &= (d-1, 0), \\ (\alpha_{\mathrm{S}}, \tau_{\mathrm{S}}) &= \left(\frac{d}{q} - 1 + \frac{d - d/q}{m}, \frac{d - d/q}{m} \right), \\ (\alpha_{\mathrm{I}}, \tau_{\mathrm{I}}) &= \left(\frac{d-1}{m}, \frac{d-1}{m} \right). \end{aligned} \tag{161}$$

In particular, $V_s + U_s = (d-1)\|\bar{a}\|^2 + dH_M^2$ is independent of the coupling.

Proof. The sparse Rademacher identities give

$$\mathbb{E} D_e = I, \qquad D_e^2 = d D_e, \qquad \mathbb{E}\|D_e a\|^2 = d\|a\|^2, \qquad \mathbb{E}\|(D_e - I)a\|^2 = (d-1)\|a\|^2. \tag{162}$$

Under G, $\bar{v} = D\bar{a}$, so $V_{\mathrm{G}} = (d-1)\|\bar{a}\|^2$. Under I, centered pair estimates are independent. Consequently

$$V_{\mathrm{I}} = \frac{d-1}{m^2} \sum_e \|a_e\|^2 = \frac{d-1}{m} \big(\|\bar{a}\|^2 + H_M^2 \big). \tag{163}$$

For S, condition first on the common projector P. Independence of the signs gives $\mathbb{E}[D_e \mid P] = (d/q)P$. The support component of the variance is

$$\mathbb{E}\|((d/q)P - I)\bar{a}\|^2 = (d/q - 1)\|\bar{a}\|^2. \tag{164}$$

The conditional sign fluctuations are independent across pairs and centered given P. Their averaged contribution is

$$\frac{1}{m^2} \sum_e \mathbb{E}\|(D_e - (d/q)P)a_e\|^2 = \frac{d - d/q}{m^2} \sum_e \|a_e\|^2 \tag{165}$$

$$= \frac{d - d/q}{m} \big(\|\bar{a}\|^2 + H_M^2 \big). \tag{166}$$

The law of total variance proves the middle line of (161). Finally, the pathwise identity

$$\frac{1}{m} \sum_e \|v_e - \bar{v}\|^2 = \frac{1}{m} \sum_e \|v_e\|^2 - \|\bar{v}\|^2 \tag{167}$$

and (162) give $U_s = d(\|\bar{a}\|^2 + H_M^2) - \|\bar{a}\|^2 - V_s$. □

For $d > 1$ and $m > 1$, I has lower average-estimation error than G exactly when

$$H_M^2 < (m-1)\|\bar{a}\|^2. \tag{168}$$

Any decrease in V_s is an equal increase in U_s. Thus a smaller average-estimator variance does not alone order the optimization trajectories. At $q = 1$, G and S coincide at the level of D_e because the sign cancels. At $q = d$, S and I have the same operator law because the support is fixed. For $N = 2$, all variants have the same law because there is only one pair. If $d = 1$, every $D_e = 1$ and all V_s vanish.

B.3 The Coupled Mixing Step

Stack states as rows of X and let $\Pi = I - N^{-1}\mathbf{1}\mathbf{1}^\top$. With $L_e = (e_i - e_j)(e_i - e_j)^\top$, the mixing operator is

$$\mathcal{A}X = X - \frac{1}{2} \sum_{e \in M} L_e X P_e. \tag{169}$$

Disjoint edges imply that $\mathcal{A}$ is an orthogonal projection in the Frobenius inner product. It preserves the average and commutes with Π.

Let V have row v_e at both endpoints of e. The actual ideal-direction update is

$$X^+ = \mathcal{A}X - \eta V. \tag{170}$$

This follows by mixing the two local estimates: averaging $D_e \nabla F_i(x_i)$ and $D_e \nabla F_j(x_j)$ gives $D_e a_e$. Because the two rows of V agree within every pair,

$$\mathcal{A}V = V, \qquad \langle \Pi \mathcal{A} X, \Pi V \rangle_F = \langle \Pi X, \Pi V \rangle_F. \tag{171}$$

No independence between V and the query–mixing mask is used here.

Proposition 4. *For fixed X, M, define $\bar{x} = N^{-1} \sum_i x_i$, $\mathcal{D} = N^{-1} \|\Pi X\|_F^2$, and $C_M = N^{-1} \sum_{e=\{i,j\}} \langle x_i + x_j - 2\bar{x}, a_e \rangle$. Under the ideal-direction update,*

$$\mathbb{E}\mathcal{D}^+ = \mathcal{D} - \frac{p}{2N} \sum_{\{i,j\} \in M} \|x_i - x_j\|^2 - 2\eta C_M + \eta^2 U_s. \tag{172}$$

Proof. Orthogonality gives the exact pure-mixing energy drop

$$\|\Pi \mathcal{A} X\|_F^2 = \|\Pi X\|_F^2 - \frac{1}{2} \sum_e \|P_e(x_i - x_j)\|^2 .$$

Take its support expectation, expand (170), and apply (171). Since $\mathbb{E} v_e = a_e$, the cross term is C_M. The normalized update energy is precisely U_s. □

Equation (172) separates pure mixing, gradient alignment, and update disagreement. Pair constancy determines the cross term while retaining the shared query–mixing mask. Proposition 7 gives the moments with stored momentum, and Theorem 5 proves its convergence.

B.4 Graph and Objective Assumptions

Assumption 5. *Each deterministic objective F_i is L-smooth, $L > 0$, and $F = N^{-1} \sum_i F_i$ satisfies $F \geq F^* > -\infty$. For all x,*

$$\frac{1}{N} \sum_i \|\nabla F_i(x) - \nabla F(x)\|^2 \leq \zeta^2. \tag{173}$$

Assumption 6. *For every used direction and point, the central-difference remainder obeys*

$$|F_i(x + \mu u) - F_i(x - \mu u) - 2\mu \nabla F_i(x)^\top u| \leq (L_3/3)\mu^3 \|u\|^3 . \tag{174}$$

Here the objectives are deterministic and nonconvex, with uniformly bounded gradient heterogeneity and unrestricted absolute gradients. Theorem 4 instead uses the stochastic-oracle assumptions from Section 4.

Matchings are independent draws from a fixed state-independent distribution on perfect matchings. Write

$$\bar{L} = \mathbb{E}_M \sum_{e \in M} L_e, \qquad \chi = 1 - \frac{\lambda_2(\bar{L})}{2} < 1, \qquad \gamma = \frac{p \lambda_2(\bar{L})}{2} = p(1 - \chi) > 0. \tag{175}$$

Because matching Laplacians have spectrum in $[0, 2]$, $0 \leq \chi < 1$. With $\mathcal{F}_t$ containing only past randomness, X_t is fixed when the current matching and directions are sampled. The pure-mixing contraction therefore applies to X_t, not to a mask-dependent post-query state.

There is a second role for the matching spectrum. Let G stack the current local gradients and $\bar{a}$ be their average. Pair averaging all coordinates uses the projection $T_M = I - \frac{1}{2} \sum_e L_e$. Hence

$$\begin{aligned} \mathbb{E}_M H_M^2 &= \frac{1}{N} \operatorname{tr}\big((\Pi G)^\top (I - \bar{L}/2) \Pi G\big) \\ &\leq \frac{\chi}{N} \sum_i \|\nabla F_i(x_i) - \bar{a}\|^2 \leq 2\chi(L^2 \mathcal{D} + \zeta^2). \end{aligned} \tag{176}$$

The last inequality uses the network gradient as a candidate center, smoothness, and (173). Also

$$\|\bar{a} - \nabla F(\bar{x})\|^2 \leq L^2 \mathcal{D}, \qquad \|\bar{a}\|^2 \leq 2 \|\nabla F(\bar{x})\|^2 + 2L^2 \mathcal{D}. \tag{177}$$

On complete graphs, uniform perfect matchings give $\chi = (N-2)/(2(N-1))$. On even rings with independently sampled matching phases, $\chi = (1+\cos(2\pi/N))/2$. The Qwen ring uses deterministic alternation, whose two-round contraction is proved in Appendix A.20.

B.5 A Correlation-Dependent Core Rate

For a fixed coupling s, abbreviate $\alpha = \alpha_s$, $\tau = \tau_s$ and define

$$r = d - 1 - \alpha, \quad J = \chi(d - \tau + 2/\gamma), \quad K = r + J, \quad B_\mu = \frac{L_3 \mu^2 dq}{6}. \tag{178}$$

Theorem 6. *Under Assumptions 5–6 and* (175)*, run the two-query matching core with common initialization, fixed coupling, and constant $\eta, \mu > 0$. If*

$$\eta \le \min\left\{ \frac{1}{8L(1+\alpha+\tau\chi)}, \frac{\sqrt{\gamma}}{16L\sqrt{K}} \right\}, \tag{179}$$

where the second restriction is omitted for $K = 0$, then with $\Delta_0 = F(\bar{x}_0) - F^*$,

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\left\|\nabla F(\bar{x}_t)\right\|^2 \le \mathcal{O}\Bigg(\frac{\Delta_0}{\eta T} + L\eta\tau\chi\zeta^2 + \frac{\eta^2 L^2 J}{\gamma}\zeta^2 + \left[1 + \frac{\eta^2 L^2}{\gamma}(1+\gamma^{-1})\right] B_\mu^2 \Bigg). \tag{180}$$

For fixed problem and graph parameters, $\eta = \Theta(T^{-1/2})$ and $\mu = \Theta((dq)^{-1/2}T^{-1/4})$ give an $O(T^{-1/2})$ bound. G has $\tau = 0$: taking $\eta = \Theta(T^{-1/3})$ and $\mu = \Theta((dq)^{-1/2}T^{-1/6})$ yields $O(T^{-2/3})$. The step condition (179) holds for sufficiently large horizons. These are upper bounds; no matching lower bound is established for S or I.

For a nominal per-node value-payload budget $\mathcal{B}$, substitute $T = \lfloor \mathcal{B}/(qB_{\mathrm{val}}) \rfloor$ whenever $T \ge 1$. The couplings have identical payloads and round counts, so (180) isolates their optimization differences. The bound uses real-valued updates, omits quantization error, and retains dependence on $d, q, N, L, \gamma, \chi, \zeta$.

B.5.1 Proof of Theorem 6

Condition on $\mathcal{F}_t$ and suppress t. Taylor's bound gives

$$\hat{g}_i = D_e \nabla F_i(x_i) + e_i, \qquad \|e_i\| \le B_\mu. \tag{181}$$

After pair averaging the applied update has identical rows $D_e a_e + r_e$ at the two endpoints, where $r_e = (e_i + e_j)/2$ and $\|r_e\| \le B_\mu$. Thus

$$X^+ = \mathcal{A}X - \eta(V + R), \quad \mathcal{A}(V+R) = V + R, \quad \bar{x}^+ = \bar{x} - \eta(\bar{v} + \bar{r}). \tag{182}$$

The pair-invariance identity holds for the residual as well, even though it depends on the direction.

Let $g = \nabla F(\bar{x})$. Equations (177) and (181) imply

$$\mathbb{E}\langle g, \bar{v} + \bar{r}\rangle \ge \frac{3}{4}\left\|g\right\|^2 - 2L^2\mathcal{D} - 2B_\mu^2. \tag{183}$$

Indeed the conditional mean error from g has squared norm at most $2L^2\mathcal{D} + 2B_\mu^2$, and $\langle g, e\rangle \ge -\left\|g\right\|^2/4 - \left\|e\right\|^2$. By Proposition 1, $\|\bar{r}\| \le B_\mu$, and (176),

$$\mathbb{E}\left\|\bar{v} + \bar{r}\right\|^2 \le 4(1+\alpha)\left\|g\right\|^2 + 4L^2(1+\alpha+\tau\chi)\mathcal{D} + 4\tau\chi\zeta^2 + 2B_\mu^2. \tag{184}$$

Smooth descent and the first condition in (179) now give

$$\mathbb{E}F(\bar{x}^+) \le F(\bar{x}) - \frac{\eta}{2}\left\|g\right\|^2 + \frac{9}{4}\eta L^2\mathcal{D} + 2L\eta^2\tau\chi\zeta^2 + 3\eta B_\mu^2. \tag{185}$$

Here $L\eta \le 1/8$, the gradient coefficient is at most $-3\eta/4 + 2L\eta^2(1+\alpha) \le -\eta/2$, and the disagreement coefficient is at most $2\eta L^2(1+1/8) = 9\eta L^2/4$.

For disagreement, expand the squared norm of the actual update. Apply pure-mixing contraction only to X, then apply pair invariance to the cross term. The ideal cross term has magnitude at most

$\sqrt{\mathcal{D}}\sqrt{\mathbb{E}_M H_M^2}$; the residual contributes at most $B_\mu\sqrt{\mathcal{D}}$. Young's inequality separately bounds each term by $\gamma\mathcal{D}/4$ plus $4\eta^2/\gamma$ times its corresponding squared magnitude. Also $\mathbb{E}\,\|\Pi(V+R)\|_F^2/N \le 2\mathbb{E}_M U_s + 2B_\mu^2$. Consequently,

$$\begin{aligned}\mathbb{E}\mathcal{D}^+ &\le (1-\gamma/2)\mathcal{D} + 2\eta^2 r\,\|\bar{a}\|^2 + 2\eta^2(d-\tau+2/\gamma)\mathbb{E}_M H_M^2 + \eta^2(2+4/\gamma)B_\mu^2 \\ &\le (1-\gamma/2+4\eta^2L^2K)\mathcal{D} + 4\eta^2 r\,\|g\|^2 + 4\eta^2 J\zeta^2 + \eta^2(2+4/\gamma)B_\mu^2. \end{aligned} \tag{186}$$

The second condition in (179) implies $\eta^2L^2K \le \gamma/256$; in particular the coefficient on $\mathcal{D}$ is at most $1-\gamma/4$.

Use the potential

$$\Phi_t = F(\bar{x}_t) + \frac{12\eta L^2}{\gamma}\mathcal{D}_t. \tag{187}$$

Adding the weighted form of (186) to (185) absorbs the disagreement term. Since $r \le K$, $48\eta^3L^2r/\gamma \le 3\eta/16 \le \eta/4$. Dropping the remaining negative disagreement term gives

$$\begin{aligned}\mathbb{E}[\Phi_{t+1} \mid \mathcal{F}_t] \le \Phi_t - \frac{\eta}{4}\,\|\nabla F(\bar{x}_t)\|^2 + 2L\eta^2\tau\chi\zeta^2 + \frac{48\eta^3L^2J}{\gamma}\zeta^2 \\ + \left[3\eta + \frac{12\eta^3L^2}{\gamma}(2+4/\gamma)\right]B_\mu^2. \end{aligned} \tag{188}$$

Take expectations, sum over $t<T$, and use $\Phi_0 = F(\bar{x}_0)$ and $\Phi_T \ge F^*$. Division by $\eta T/4$ gives the explicit bound

$$\begin{aligned}\frac{1}{T}\sum_{t<T}\mathbb{E}\,\|\nabla F(\bar{x}_t)\|^2 \le \frac{4\Delta_0}{\eta T} + 8L\eta\tau\chi\zeta^2 + \frac{192\eta^2L^2J}{\gamma}\zeta^2 \\ + \left[12 + \frac{48\eta^2L^2}{\gamma}\left(2+\frac{4}{\gamma}\right)\right]B_\mu^2. \end{aligned} \tag{189}$$

This proves (180). No uniform bound on $\|\nabla F_i(x_i)\|$ was used. □

B.6 A Quadratic Cancellation Example

The trajectory experiment uses

$$F_i(x) = \tfrac{1}{2}x^\top Hx + b_i^\top x, \qquad H = \mathrm{diag}(h_1,\dots,h_d) \succ 0, \qquad \sum_i b_i = 0. \tag{190}$$

Then $F(x) = x^\top Hx/2$ and $\zeta^2 = N^{-1}\sum_i \|b_i\|^2$. Central differences are exact for every $\mu > 0$ in ideal arithmetic.

Proposition 5. *For G on* (190)*, the network average satisfies*

$$\bar{x}_{t+1} = \bar{x}_t - \eta D_t H\bar{x}_t \tag{191}$$

regardless of the matching sequence and the individual linear terms. At the consensus optimum $x_i = 0$, a single step of coupling s instead has $\mathbb{E}\,\|\bar{x}^+\|^2 = \eta^2\tau_s\mathbb{E}_M H_M^2$.

Proof. Average preservation and a common D_t give $N^{-1}\sum_i \hat{g}_i = D_t(H\bar{x}_t + N^{-1}\sum_i b_i) = D_tH\bar{x}_t$. At consensus zero, $\bar{a} = 0$, so Proposition 1 gives the second assertion. □

Shared curvature makes the G average trajectory identical across these heterogeneity levels and topologies. Proposition 6 below derives the residual when local Hessians differ, and Proposition 7 adds oracle noise and momentum.

B.7 Mechanism Experiments

The configuration was fixed before training and archived with a timestamp and source hashes. There are $N = 8$ nodes, $d = 32$, $q = 4$, eight seeds, three couplings, two topologies, three heterogeneity levels, and two step sizes: 288 runs. Each uses 1600 two-query rounds and 3200 optimization queries

per node. The bit axis charges a nominal 32 bits per value, or 204800 bits per node. Computation uses NumPy double precision without packet quantization or physical transfer. All runs use the deterministic core, fixed parameters, and unclipped updates.

The eigenvalues of H in (190) range geometrically from 1 to 4. Gaussian b_i are centered and scaled to $\zeta \in \{0, 0.5, 2\}$, and all nodes start from $d^{-1/2}\mathbf{1}$. Public directions and matchings are generated independently of these linear terms. Ring phases are independent fair draws at each round, rather than deterministic alternation. Within every comparison, supports/signs are the only protocol change. Objective and gradient diagnostics are evaluation-only and do not enter the update.

The step $\eta = 10^{-4}$ satisfies (179) for every reported coupling and topology. The larger $\eta = 0.004$ tests behavior outside that sufficient condition. All runs at both steps are finite and included in the results.

For the frozen-state study, $\|\bar{a}\| = 1$ and H_M^2 takes six specified values. Each seed averages 1500 direction draws for each coupling; gradients are held fixed, so these are random-matrix moment diagnostics rather than optimization-query comparisons. Figure 5 overlays the exact formulas and measured means. At $H_M^2 = 0$, the predicted G/S/I average variances are 31, 13, and 7.75; at $H_M^2 = 15$ they are 31, 103, and 124. All three intersect at $H_M^2 = m-1 = 3$, with the complementary ordering in disagreement injection.

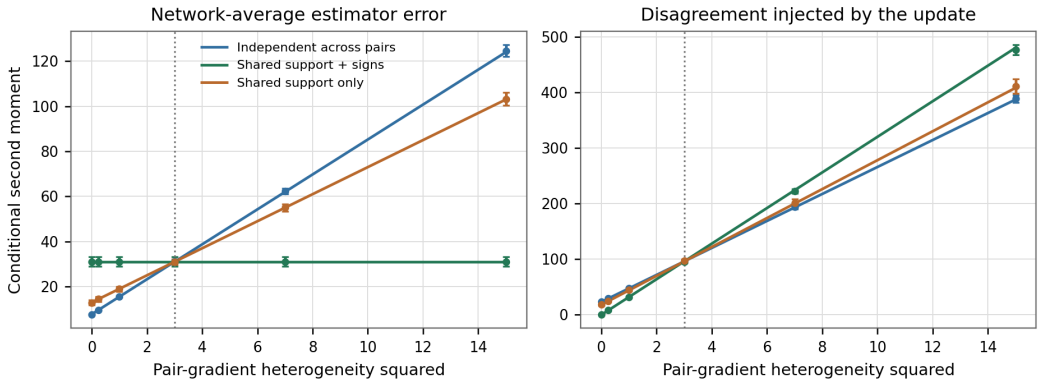


Figure 5: Exact moment curves and eight-seed Monte Carlo means ± standard deviation. The dotted line is the predicted crossover.

At $\eta = 0.004$, nonzero linear heterogeneity produces an optimization floor for S and I, while G retains the cancellation in (191). For $\zeta = 2$ on the complete graph, the final objective is $(1.160 \pm 0.421) \times 10^{-7}$ for G and 0.01118 ± 0.00309 for I. At zero heterogeneity, the final means are close despite the different frozen-state variances. The separation in this common-curvature example is driven by heterogeneity cancellation.

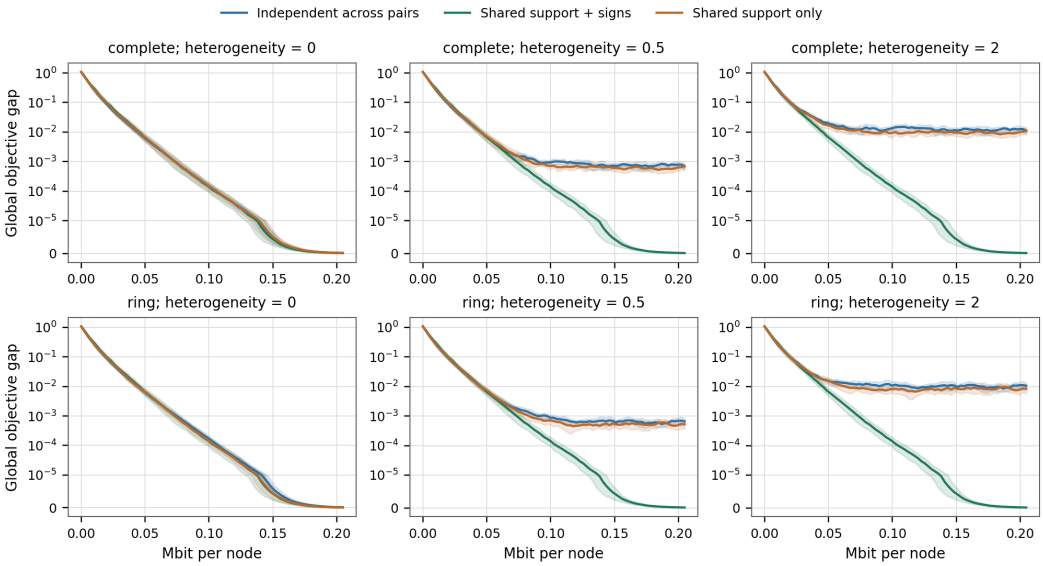


Figure 6: Objective gap at $\eta = 0.004$ (outside the sufficient step certificate). Curves are eight-seed means and bands are one standard deviation.

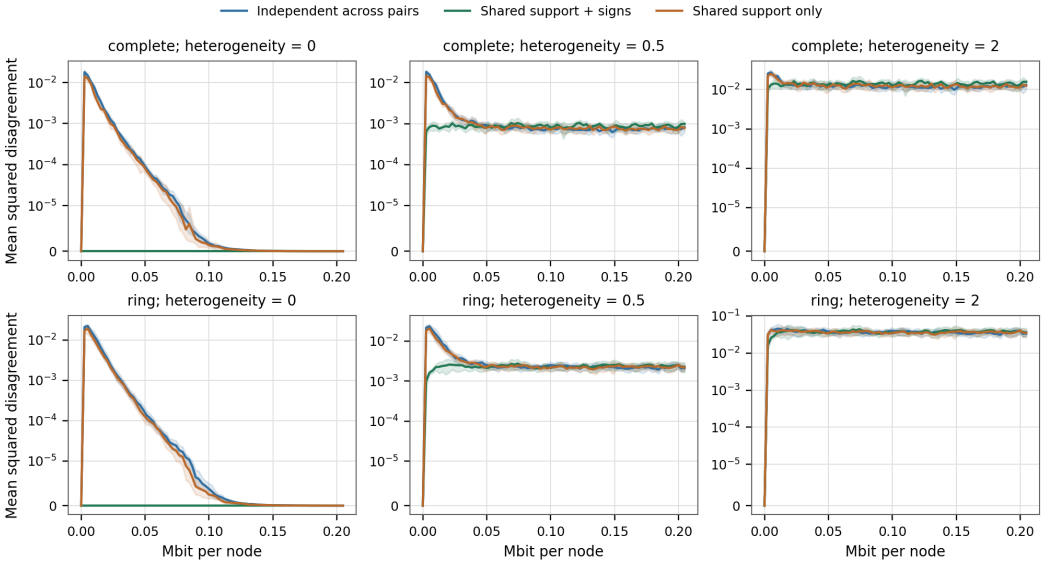


Figure 7: State disagreement for the same eight-seed runs.

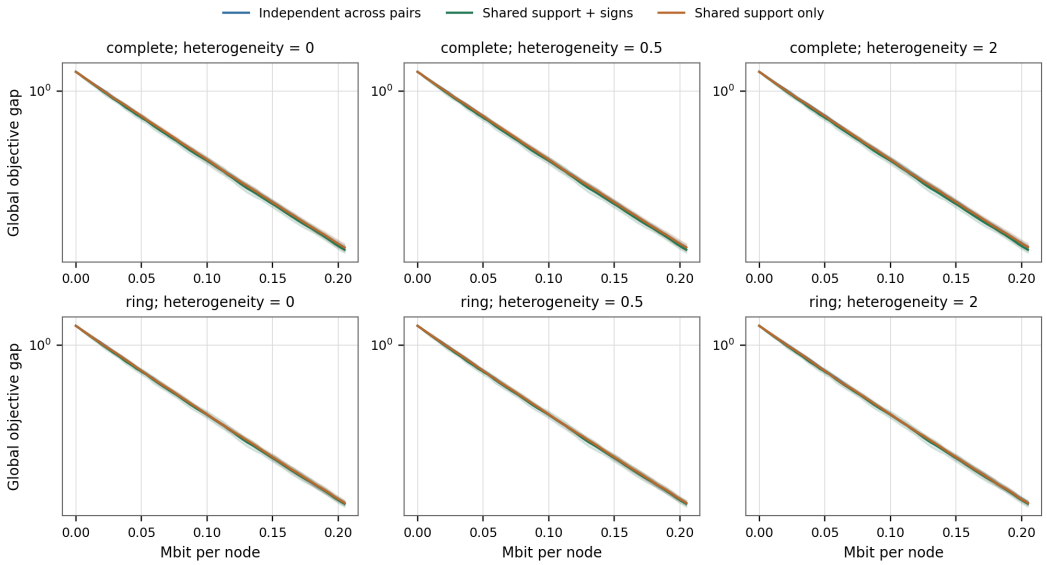


Figure 8: Objective gap at $\eta = 10^{-4}$, which satisfies the sufficient step condition for all displayed settings.

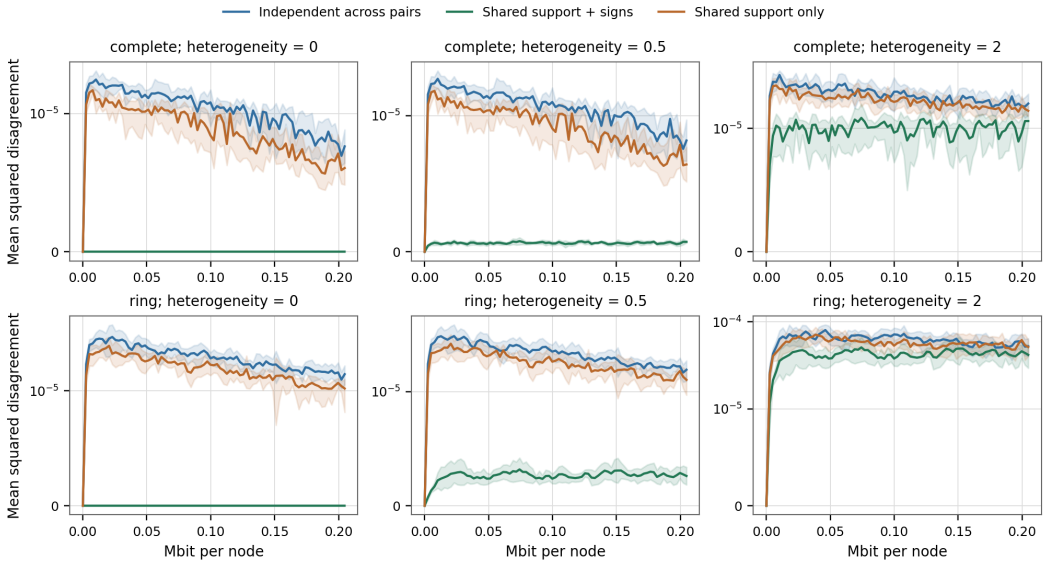


Figure 9: State disagreement at the certified step. Training plots use a symmetric-log scale, linear below 10^{-5}; standard-deviation bands are truncated at zero for nonnegative metrics.

B.8 Curvature Heterogeneity

The common-Hessian example isolates exact cancellation. With different local curvatures, the uncancelled part can be identified rather than absorbed into an absolute gradient bound. Let

$$F_i(x) = \tfrac{1}{2}x^\top H_i x + b_i^\top x, \qquad H = \frac{1}{N}\sum_i H_i, \quad b = \frac{1}{N}\sum_i b_i, \quad \delta_H^2 = \frac{1}{N}\sum_i \|H_i - H\|_{\mathrm{op}}^2,$$

where the Hessians are symmetric. Write $\mathcal{D} = N^{-1}\sum_i \|x_i - \bar{x}\|^2$.

Proposition 6. *For arbitrary node states,*

$$\bar{a} = \nabla F(\bar{x}) + r_X, \qquad r_X = \frac{1}{N}\sum_i (H_i - H)(x_i - \bar{x}), \qquad \|r_X\|^2 \le \delta_H^2 \mathcal{D}. \tag{192}$$

Under G without momentum or oracle noise, the network-average step is $\bar{x}^+ = \bar{x} - \eta D(\nabla F(\bar{x}) + r_X)$ and

$$\mathbb{E}\|D r_X\|^2 = d\|r_X\|^2 \le d\delta_H^2 \mathcal{D}. \tag{193}$$

Proof. Expand $H_i x_i$ around $\bar{x}$. Since $\sum_i (x_i - \bar{x}) = 0$,

$$\frac{1}{N}\sum_i (H_i x_i + b_i) = H\bar{x} + b + \frac{1}{N}\sum_i (H_i - H)(x_i - \bar{x}).$$

The triangle and Cauchy–Schwarz inequalities give

$$\|r_X\| \le \frac{1}{N}\sum_i \|H_i - H\|_{\mathrm{op}}\|x_i - \bar{x}\| \le \delta_H\sqrt{\mathcal{D}}.$$

The common direction commutes with averaging, and $\mathbb{E}D^2 = dI$ gives (193). □

The averaging identity holds with different Hessians, with a residual controlled by state disagreement. Shared curvature or consensus gives $r_X = 0$. This identity applies to arbitrary quadratic states; applying Theorem 6 additionally requires its uniform gradient-heterogeneity assumption over $\mathbb{R}^d$.

B.9 Oracle Noise and Stored Momentum

Condition on the node states, stored local momenta, and current matching before drawing directions and oracle samples. For $e = \{i, j\}$, retain a_e from (157) and put $c_e = (m_i + m_j)/2$, $\bar{c} = m^{-1}\sum_e c_e$, and $\theta = 1 - \beta$, with $0 \le \beta < 1$. Suppose local gradient errors ϵ_i are centered, mutually independent, and independent of the current directions under this conditioning. Define

$$\sigma_X^2 = \frac{1}{N}\sum_i \mathbb{E}\|\epsilon_i\|^2, \qquad \epsilon_e = (\epsilon_i + \epsilon_j)/2, \qquad p = q/d.$$

Neither equal node variances nor isotropic noise is needed. The ideal applied pair update and its network-average conditional mean are

$$w_e = \beta P_e c_e + \theta D_e(a_e + \epsilon_e), \qquad \nu = \mathbb{E}\bar{w} = \beta p\bar{c} + \theta\bar{a}. \tag{194}$$

This is the sparse-momentum rule (3) after mixing. Inactive memory entries are retained, not decayed.

For $a, c \in \mathbb{R}^d$, define the nonnegative quadratic form

$$\Psi_\beta(a, c) = \beta^2 p(1-p)\|c\|^2 + \theta^2(d-1)\|a\|^2 + 2\beta\theta(1-p)\langle c, a\rangle. \tag{195}$$

The cross term is essential: the stored direction and fresh query use the same support.

Proposition 7. *With $V_s^{\beta,\sigma} = \mathbb{E}\|\bar{w} - \nu\|^2$, the three couplings satisfy*

$$V_{\mathrm{G}}^{\beta,\sigma} = \Psi_\beta(\bar{a}, \bar{c}) + \theta^2 d\sigma_X^2/N, \tag{196}$$

$$V_{\mathrm{I}}^{\beta,\sigma} = \frac{1}{m^2}\sum_e \Psi_\beta(a_e, c_e) + \theta^2 d\sigma_X^2/N, \tag{197}$$

$$V_{\mathrm{S}}^{\beta,\sigma} = p(1-p)\|\beta\bar{c} + \theta\bar{a}/p\|^2 + \frac{\theta^2(d - 1/p)}{m^2}\sum_e \|a_e\|^2 + \theta^2 d\sigma_X^2/N. \tag{198}$$

For $U_s^{\beta,\sigma} = m^{-1}\sum_e \mathbb{E}\|w_e - \bar{w}\|^2$,

$$V_s^{\beta,\sigma} + U_s^{\beta,\sigma} = \frac{1}{m}\sum_e \left[\beta^2 p\|c_e\|^2 + \theta^2 d\|a_e\|^2 + 2\beta\theta\langle c_e, a_e\rangle\right] + \frac{\theta^2 d\sigma_X^2}{2} - \|\nu\|^2. \tag{199}$$

Proof. Because $PD = D$, $\mathbb{E}P = pI$, and $\mathbb{E}D = I$, we have

$$\mathbb{E}(P - pI)^2 = p(1-p)I, \qquad \mathbb{E}(D - I)^2 = (d-1)I, \qquad \mathbb{E}[(P - pI)(D - I)] = (1-p)I.$$

Expanding the squared norm of $\beta(P - pI)c + \theta(D - I)a$ proves (195) and its nonnegativity. For G, both random maps are common, so the noiseless average is $\beta P\bar{c} + \theta D\bar{a}$. For I, centered pair updates are independent; the variance of their average is the sum of variances divided by m^2. These observations give the noiseless parts of (196)–(197).

For S, condition first on the common P. The conditional mean of the noiseless average is $P(\beta\bar{c} + \theta\bar{a}/p)$. Its variance is the first term of (198). Conditional sign fluctuations are independent across pairs. Their variance is $\theta^2(d - 1/p)m^{-2}\sum_e \|a_e\|^2$, by the same conditional calculation as in Appendix B.2.

Oracle cross terms vanish when conditioned on all directions. In every coupling,

$$\mathbb{E}\left\|\frac{\theta}{N}\sum_i D_{e(i)}\epsilon_i\right\|^2 = \frac{\theta^2}{N^2}\sum_i \mathbb{E}[\epsilon_i^\top D_{e(i)}^2\epsilon_i] = \frac{\theta^2 d\sigma_X^2}{N}.$$

The first equality uses independence across nodes even when their directions coincide; the second uses independence from directions and $\mathbb{E}D_e^2 = dI$. The oracle component is also orthogonal in expectation to the noiseless update.

Finally,

$$\frac{1}{m}\sum_e \mathbb{E}\|w_e\|^2 = \frac{1}{m}\sum_e [\beta^2 p\|c_e\|^2 + \theta^2 d\|a_e\|^2 + 2\beta\theta\langle c_e, a_e\rangle] + \frac{\theta^2 d\sigma_X^2}{2}.$$

Subtracting $\|\nu\|^2$ and using the pathwise average/disagreement identity proves (199). □

In particular, at $\beta = 0$, oracle noise adds $d\sigma_X^2/N$ to each average variance and $d\sigma_X^2(1/2 - 1/N)$ to each injection term. Sharing directions preserves cancellation of deterministic heterogeneity, not cancellation of independent oracle errors.

The momentum comparison is equally explicit. If $\mathcal{H}_\Psi = m^{-1}\sum_e \Psi_\beta(a_e - \bar{a}, c_e - \bar{c})$, then

$$V_{\mathrm{I}}^{\beta,\sigma} - V_{\mathrm{G}}^{\beta,\sigma} = \frac{\mathcal{H}_\Psi - (m-1)\Psi_\beta(\bar{a}, \bar{c})}{m}. \tag{200}$$

The comparison conditions on the same states and stored memories; both evolve differently along separate training trajectories.

Finite differences introduce a controlled perturbation. Let $\widetilde{w}_e$ be the actual applied pair update under Assumption 4, with the same oracle sample used in the two queries. Let $\widetilde{V}_s = \mathbb{E}\|\widetilde{\bar{w}} - \mathbb{E}\widetilde{\bar{w}}\|^2$. Then

$$\|\mathbb{E}\widetilde{\bar{w}} - \nu\| \le \theta B_\mu, \qquad \left|\sqrt{\widetilde{V}_s} - \sqrt{V_s^{\beta,\sigma}}\right| \le \theta B_\mu. \tag{201}$$

Indeed, Taylor's remainder bounds each local estimation error by B_μ. Pair and network averaging therefore bound the extra applied vector by θB_μ. Centering cannot increase its second moment, and the reverse triangle inequality in L^2 gives the second inequality. Stacking the actual applied pair vectors as rows of $\widetilde{W}$, pair constancy gives $\mathcal{A}\widetilde{W} = \widetilde{W}$ also holds.

These formulas give the conditional moments for noisy sparse momentum with finite differences. Theorem 5 supplies the corresponding convergence guarantee.

B.10 Extended Mechanism Experiments

The follow-up study changes local curvature, oracle noise, and sparse momentum while preserving the matching and payload comparison. Its configuration and main-panel selection were fixed before

training and archived with a timestamp and source hashes in `results/correlation_v23/protocol.json`. All 1536 runs are retained.

There are eight nodes and $d = 32$ coordinates, with $q = 4$, $\mu = 10^{-3}$, and 1600 rounds. Write $h = \mathrm{geomspace}(1, 4, d)$. For each seed, a Gaussian matrix is centered by coordinate and divided by its largest absolute entry to obtain R, with $N^{-1}\sum_i R_i = 0$ and $|R_{ij}| \le 1$. We set

$$H_i = \mathrm{diag}(h_j(1 + \omega R_{ij}))_{j=1}^{d}, \qquad \omega \in \{0, 0.6\}.$$

Thus all Hessians are positive, their average is fixed, and $\omega = 0.6$ introduces local curvature differences without changing the global objective. Gaussian linear terms are centered and scaled to $N^{-1}\sum_i \|b_i\|^2 = \zeta^2$, with $\zeta \in \{0, 2\}$. Here ζ describes the linear terms, not a uniform gradient-heterogeneity bound for unequal Hessians.

The oracle is $f_i(x; \epsilon_i) = F_i(x) + \epsilon_i^\top x$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_d/d)$. Samples are independent across nodes and rounds and reused within the two-point difference. They are paired across the compared couplings and parameters. The updates use the two function values, not analytic gradients; analytic quantities are recorded only for evaluation. Central differences are exact for these quadratic-plus-linear samples up to floating-point error.

For each coupling, we test complete-graph uniform perfect matchings and independently sampled ring phases, eight seeds, both curvature settings, both linear-heterogeneity settings, and base steps $\eta_0 \in \{0.001, 0.004\}$. Four update settings complete the grid: deterministic core $(\beta, \sigma, \eta) = (0, 0, \eta_0)$; noisy core $(0, 0.5, \eta_0)$; noisy sparse momentum $(0.9, 0.5, \eta_0)$; and first-step-matched noisy momentum $(0.9, 0.5, 10\eta_0)$. The last choice makes $(1 - \beta)\eta = \eta_0$, so its first update agrees with the noisy core for each realization when memory starts at zero. Subsequent updates need not agree.

Every run uses 3200 optimization queries and a nominal 204800 outgoing bits per node. Computation is float64; values are charged at 32 bits without quantizing or transmitting packets. All runs remained finite, with no clipping, early stopping, or seed replacement. The measured residual satisfied (192) within numerical precision at every logged state.

For $\omega = 0.6$, $\zeta = 2$, and $\eta_0 = 0.004$, G remains below S and I at the same step under deterministic, noisy, and sparse-momentum updates (Figures 10–11). On the complete graph, the deterministic final objective is $(2.96 \pm 1.13) \times 10^{-5}$ for G and $(1.34 \pm 0.48) \times 10^{-2}$ for I. On the ring the G residual is larger, $(1.43 \pm 0.49) \times 10^{-4}$, consistent with the loss of topology-independent cancellation when curvatures differ.

The ordering also depends on step size. In the first-step-matched control at $\eta = 0.04$, G has a much larger final objective than I on both graphs. At the smaller base step, its first-step-matched $\eta = 0.01$ retains the advantage when $\zeta = 2$. This dependence is compatible with (200), which compares conditional moments at a common state and memory, rather than complete trajectories.

When $\zeta = 0$, the large deterministic advantage disappears (Figures 12–13), identifying heterogeneity as the source of the separation. Oracle noise raises the floor reached by G above its deterministic value.

The eight-seed means and sample standard deviations for every configuration are in `results/correlation_v23/final_summary.csv`, alongside raw trajectories, per-seed endpoints, and a validation audit. The archive contains objective and state-disagreement plots for all 16 combinations of graph, curvature, linear heterogeneity, and base step. The figures below include both momentum step-size choices.

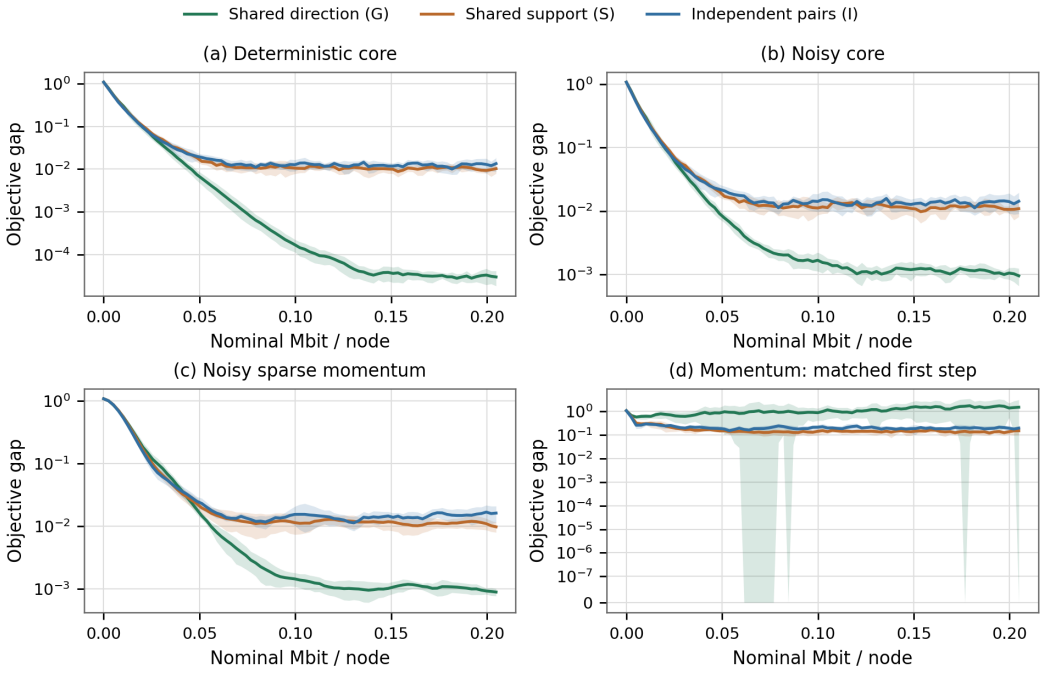


Figure 10: Unequal curvatures, $\zeta = 2$, complete graph. All panels use eight seeds; (d) increases the momentum step from 0.004 to 0.04.

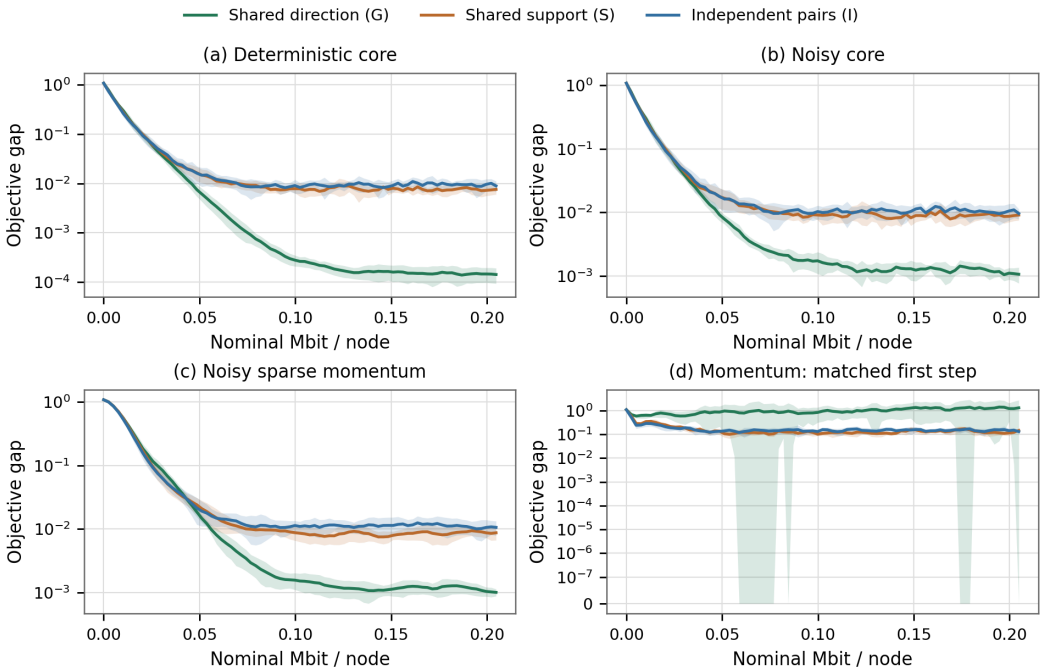


Figure 11: The same settings on iid ring matchings; (d) matches the first-update scale.

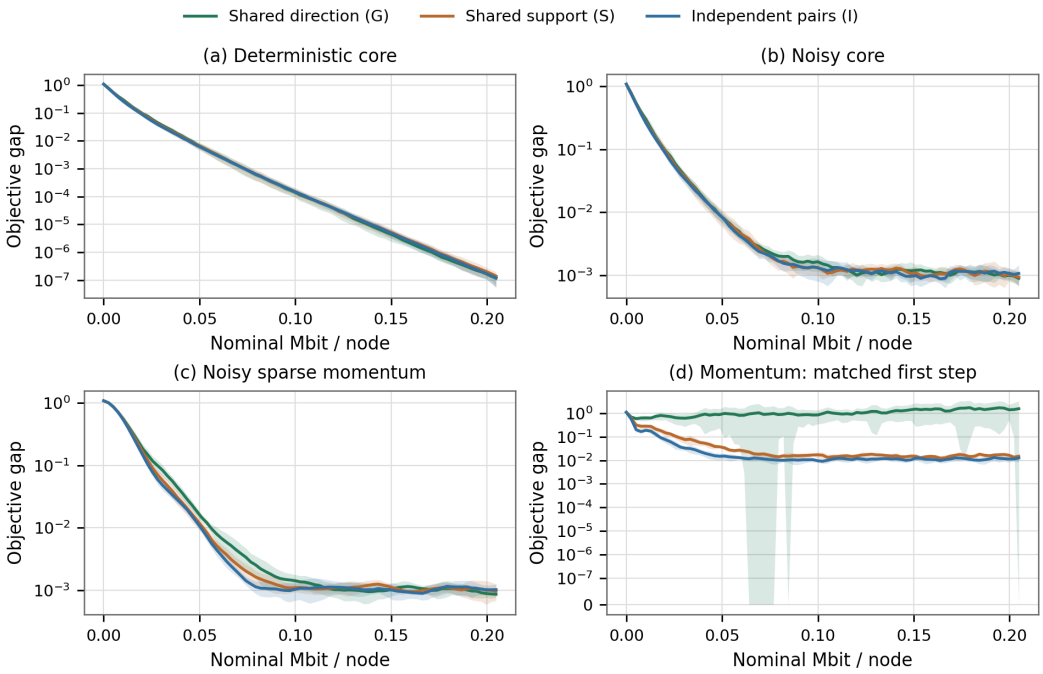


Figure 12: Unequal curvatures with zero linear heterogeneity on the complete graph. The pronounced heterogeneity-driven separation is absent.

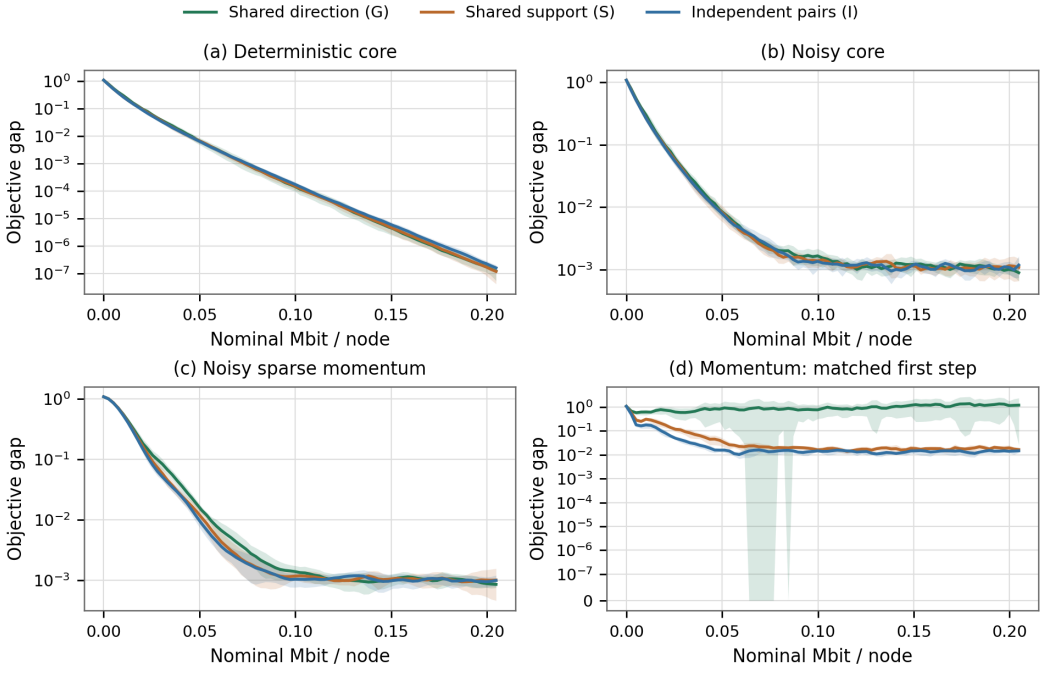


Figure 13: Zero linear heterogeneity on the ring, eight-seed mean $\pm$ std. Bands are truncated at zero for the nonnegative objective gap.

C COMMUNICATION ACCOUNTING DETAILS

An indexed sparse method sends both values and coordinate identifiers; ZO-COSMO communicates only values after the public random seed is fixed. For a q-sparse update in dimension d, let

$$\ell_d = \lceil \log_2 d \rceil \tag{202}$$

be the number of bits needed to name a coordinate. The per-round directed-link costs are

$$C_{\mathrm{COSMO}} = qB_{\mathrm{val}}, \qquad C_{\mathrm{indexed}} = q(B_{\mathrm{val}} + \ell_d). \tag{203}$$

The fraction of an indexed sparse message spent only on coordinate identities is therefore

$$r_{\mathrm{idx}}(d, B_{\mathrm{val}}) = \frac{\ell_d}{B_{\mathrm{val}} + \ell_d}. \tag{204}$$

The public-support design removes exactly this fraction of the sparse payload. The saving is modest for small dense floating-point messages, but it becomes dominant for high-dimensional models and for low-bit value quantization.

Table 2: Coordinate-identity overhead in indexed sparse communication.

Setting	Dimension d	ℓ_d	$B_{\mathrm{val}} = 32$	$B_{\mathrm{val}} = 16$	$B_{\mathrm{val}} = 8$	$B_{\mathrm{val}} = 1$
Qwen LoRA-scale experiment	2.52×10^6	22	40.7%	57.9%	73.3%	95.7%
Qwen full-model study	6.53×10^9	33	50.8%	67.3%	80.5%	97.1%

Table 2 quantifies the representation cost removed by index-free sparsity. In the Qwen full-model study, even with 32-bit values, an indexed sparse method spends slightly more than half of each sparse payload on coordinate names. If the value payload is quantized to 8 bits, the coordinate names become more than four fifths of the message. The same effect appears in LoRA-scale training, where the dimension is far smaller than full-model fine-tuning but still large enough that coordinate identifiers dominate low-bit sparse messages.

Table 3: Representative directed-link payloads.

Setting	d	q	ZO-COSMO	Private-support / refresh cost
Qwen LoRA-scale 32-bit values	2.52M	8,192	262.1K bits	$\geq$ 341.6K entropy bound; 442.4K coordinate list; 80.6M dense
Qwen full-model 32-bit values	6.53B	8,192	262.1K bits	$\geq$ 434.6K entropy bound; 532.5K coordinate list; 2.09×10^{11} dense

The entropy entries evaluate the lower bound $qB_{\mathrm{val}} + \log_2 \binom{d}{q}$ for a uniform private support. The coordinate-list entries give the implemented format. Together they distinguish the minimum support-description cost from the cost of explicit indices.

C.1 PUBLIC RANDOMNESS AND THE COMMUNICATION OPERATOR

Shared randomness can encode different mathematical objects. Table 4 separates three decentralized uses. Here $\bar{\delta}$ is average degree, D_G is graph diameter, and $b = B_{\mathrm{seed}} + B_{\mathrm{scalar}}$ is the size of one seed–scalar update. Payload is the sum of packet bits crossing edges, divided by N; sequential stages count the longest dependency chain within one optimization round.

Table 4: Payload and link stages for public-randomness protocols.

Protocol	Reconstructed object	Destination	Bits/node/round	Link stages
Global-support ZO-COSMO	Current state slice	Neighbors	$\bar{\delta}qB_{\mathrm{val}}$	1
Edge-local ZO-COSMO	Current state slice	Matched peer	qB_{val}	1
Seed-scalar flooding (Kim & Lee, 2026)	Local update	Every node	$\Omega((N-1)b)$	$\geq D_G$

Server-mediated methods send updates to a coordinator: DeComFL and MEERKAT reconstruct seed–scalar updates, while FeedSign aggregates seed–sign votes (Li et al., 2025; Ran et al., 2026; Cai

et al., 2026). Table 4 compares peer-to-peer operators, whose destinations and dependency graphs differ from this server-mediated pattern.

The flooding bound concerns delivery of every source packet to every node. Each of N distinct packets requires at least $N-1$ packet-edge deliveries, giving $N(N-1)b$ network bits, or $(N-1)b$ per node; a diameter endpoint requires at least D_G sequential links. Batching is allowed, but each original packet must reach every node; in-network aggregation is a different operator. SeedFlood uses this dissemination for all-gather-equivalent synchronization. Its SubCGE batches updates in a shared subspace, reducing application cost from $O(Nd)$ to $O(N+rd)$ for subspace rank r (Kim & Lee, 2026). The table counts packet delivery separately from this application cost.

A seed–scalar pair can be much smaller than q state values, making flooding attractive for small N or inexpensive multi-hop communication. Edge-local ZO-COSMO finishes in one link stage and keeps distinct local states, with constant per-node payload on a perfect matching. The support-locality experiment in Appendix D.2 measures the crossover at a common bit budget.

C.2 Sparse-Message Serialization Check

We measure local encoding and decoding with contiguous 32-bit floating-point values. Indexed packets additionally use `uint32` coordinates for Qwen LoRA and `uint64` for full-model Qwen. These timings cover serialization only; network transfer is measured separately below.

Table 5: Sparse-message serialization check.

Setting	Value-only	Indexed impl.	Min. indexed	Time ratio
Qwen LoRA-scale	32.0 KiB	64.0 KiB	54.0 KiB	3.22×
Qwen full-model	32.0 KiB	96.0 KiB	65.0 KiB	4.04×

The "Min. indexed" column is the ideal bit-packed cost $q(32+\ell_d)$ rounded to bytes; the indexed implementation column is the ordinary contiguous-coordinate representation used by sparse tensor code. The exact timing ratios are machine dependent, but the payload relation is not. Even under ideal bit packing, indexed sparse communication remains larger than the value-only message; under ordinary integer-coordinate serialization, the Qwen full-model indexed payload is three times the ZO-COSMO payload before any network overhead is counted.

C.3 Two-Process Message Transfer Check

We send LoRA-scale messages between two processes on the same host using PyTorch's Gloo TCP backend. The sender transmits contiguous tensors and waits for a one-byte acknowledgement. Sparse messages are repeated 300 times and the dense message 30 times within each trial; method order rotates across 31 trials. Table 6 reports median exchange time, including the acknowledgement.

Table 6: Two-process transfer of Qwen LoRA-scale messages.

Message	Tensors	Payload	Median exchange	Relative time
ZO-COSMO value-only	1	32.0 KiB	127.2 μs	1.00×
Indexed sparse	2	64.0 KiB	162.8 μs	1.28×
Dense LoRA refresh	1	9.6 MiB	2.69 ms	21.16×

Doubling the sparse payload increases median transfer time by 28%, indicating a substantial fixed-latency component. The dense refresh carries $308\times$ as many bits and takes $21.16\times$ as long on this transport. We use payload bits for the method comparisons and report these same-host timings as a separate implementation measurement.

C.4 Amortized Cost of the Shared Seed

The value-only cost in (203) assumes that the public random seed has already been fixed. It is the per-round accounting for a repeated optimization run; the one-time overhead can be written separately. Let B_{seed} be the number of bits used to initialize the public pseudorandom generator and let $B_{\mathrm{sync}}(T)$

be the total number of bits used for occasional round-counter or resynchronization control messages over T rounds. The average directed-link payload of ZO-COSMO is then

$$\bar{C}_{\mathrm{COSMO}}(T) = qB_{\mathrm{val}} + \frac{B_{\mathrm{seed}} + B_{\mathrm{sync}}(T)}{T}. \tag{205}$$

This differs from (203) only by a term that vanishes with the number of optimization rounds. By contrast, an indexed sparse method pays coordinate identifiers every round:

$$\bar{C}_{\mathrm{indexed}}(T) = qB_{\mathrm{val}} + q\ell_d. \tag{206}$$

Seed setup is amortized across rounds, whereas coordinate lists recur in every message.

For example, suppose a run uses a 128-bit seed and one 32-bit epoch counter every 100 rounds. Then

$$\frac{B_{\mathrm{seed}} + B_{\mathrm{sync}}(T)}{T} \le \frac{128}{T} + 0.32 \quad \text{bits per round.} \tag{207}$$

In the Qwen full-model setting with $q = 8192$ and 32-bit values, the value payload is $262{,}144$ bits per round and the indexed coordinate payload is $270{,}336$ bits per round. The seed and control overhead is therefore negligible relative to both the value payload and the repeated coordinate-identifier payload.

The same accounting applies to balanced public schedules. An epoch seed suffices to regenerate the support without a coordinate list. After a missed round, the round id and seed recover the support, not the missed values or optimizer state; reliable delivery or explicit state recovery is still required.

C.5 ZO-MARINA-type Baseline Definition

ZO-MARINA-type alternates sparse ZO updates with periodic dense refreshes. The name refers to this communication pattern. MARINA itself uses compressed gradient differences and a variance-reduced estimator; the comparator averages independent two-query ZO updates and has no MARINA convergence guarantee. Algorithm 3 gives the two-worker Qwen implementation.

Algorithm 3 ZO-MARINA-type for two-worker Qwen

Require: K, q, μ, η_t, β; initialize $x_{1,0} = x_{2,0}$ and $m_{1,0} = m_{2,0} = 0$

for round $t = 0, 1, \ldots, T-1$ **do**
  Set $q_t = d$ if $t \bmod K = 0$, and $q_t = q$ otherwise.
  for $i \in \{1, 2\}$ **do**
    Independently draw a size-q_t support $S_{i,t}$ and Rademacher signs, defining $u_{i,t}$.
    $g_{i,t} = \frac{d}{q_t} \frac{f_i(x_{i,t}+\mu u_{i,t};\xi_{i,t}) - f_i(x_{i,t}-\mu u_{i,t};\xi_{i,t})}{2\mu} u_{i,t}$.
    Set $P_{i,t} = \mathrm{diag}(\mathbf{1}_{S_{i,t}})$ and update $m_{i,t+1}$ by (3).
    Send the active values of $h_{i,t} = P_{i,t} m_{i,t+1}$ to the other worker.
    Include the coordinate list on sparse rounds; dense coordinates have fixed order.
  end for
  Both workers set $x_{i,t+1} = x_{i,t} - \frac{\eta_t}{2}(h_{1,t} + h_{2,t})$.
end for

On sparse rounds, each worker applies both decoded sparse updates, including overlaps. On dense rounds, it applies their full-vector average. Common initialization keeps the two replicas synchronized, so this is an update-aggregation comparison. Dense query signs are independent across workers. The LoRA experiment uses $K = 25$, with the first dense refresh at $t = 0$.

Each directed link carries $q(B_{\mathrm{val}} + \ell_d)$ bits on sparse rounds and dB_{val} bits on dense rounds. Over a complete refresh cycle, the average is

$$C_{\mathrm{MARINA}} = \frac{K-1}{K} q(B_{\mathrm{val}} + \ell_d) + \frac{1}{K} dB_{\mathrm{val}}. \tag{208}$$

For a finite horizon $T \ge 1$, there are $\lceil T/K \rceil$ dense rounds, so the cumulative cost is

$$C_T = \lceil T/K \rceil dB_{\mathrm{val}} + (T - \lceil T/K \rceil) q(B_{\mathrm{val}} + \ell_d). \tag{209}$$

With $d = 2{,}523{,}136$, $q = 8192$, and 32-bit values, the first dense packet costs 80.740 Mbit per link. Extrapolating the same packet format to full-model Qwen gives about 2.09×10^{11} bits per refresh. Seed-aware sparse protocols are evaluated separately.

Equal-round comparisons fix the query count; fixed-budget comparisons also account for the affordable horizon. When the first refresh exceeds the budget, the reported state is the initialization.

C.6 COM-DSZO BASELINE DEFINITION

Com-DSZO combines a two-point ZO estimator with error-compensated compressed consensus (Hua et al., 2026). We instantiate its published recursion with Top-k innovation compression. In matrix form, let X_t stack the node states, let $\widehat{X}_t$ be the reconstructed state, let B_t accumulate disagreement, and let Q_t be the compressed innovation. One round is

$$\widehat{X}_{t+1} = \widehat{X}_t + \psi Q_t, \tag{210}$$

$$B_{t+1} = B_t + \psi(I - W)Q_t, \tag{211}$$

$$X_{t+1} = X_t - \gamma B_{t+1} - \eta G_t^{\mathrm{ZO}}, \tag{212}$$

$$Q_{t+1} = \mathrm{Top}_q(X_{t+1} - \widehat{X}_{t+1}). \tag{213}$$

The Rosenbrock problem is unconstrained, so the projection is the identity. We use the same two function evaluations and dense Rademacher ZO estimator as the dense baselines, $\psi = 0.5$, $\gamma = 0.1$, and $\eta = 2.5 \times 10^{-3}$. Top-$q$ selects coordinates from the innovation, giving a message cost of $q(B_{\mathrm{val}} + \ell_d)$ bits. The baseline also stores dense reconstruction and disagreement arrays, $\widehat{X}_t$ and B_t. This comparison tests adaptive compression with error compensation. The separate indexed control isolates representation cost by using the ZO-COSMO trajectory and charging its coordinate list.

C.7 COMMUNICATION-BUDGET RATE

Under a directed-link budget $\mathcal{B} \geq 2qB_{\mathrm{val}}$, choose a horizon $T = \lfloor \mathcal{B}/(qB_{\mathrm{val}}) \rfloor$ satisfying Theorem 4, with its prescribed stepsize and smoothing radius. Let $R_{\mathcal{B}}^{\mathrm{COSMO}}$ denote the resulting averaged gradient norm. Then

$$R_{\mathcal{B}}^{\mathrm{COSMO}} \leq \mathcal{O}\left(\left(\Delta_0 + \frac{d\sigma^2}{N} \right) \sqrt{\frac{qB_{\mathrm{val}}}{\mathcal{B}}} + \frac{qB_{\mathrm{val}}}{\mathcal{B}} + \frac{d^2(G^2 + \sigma^2)B_{\mathrm{val}}}{(1-\rho^2)^2 \mathcal{B}} \right). \tag{214}$$

The communication-budget form of Theorem 4 follows by substituting the number of affordable rounds into the iteration bound. Let $\mathcal{B}$ denote the total number of transmitted bits per directed link. For ZO-COSMO,

$$T_{\mathrm{COSMO}} = \left\lfloor \frac{\mathcal{B}}{qB_{\mathrm{val}}} \right\rfloor. \tag{215}$$

Using common initialization and the canonical smoothing choice, the simplified bound (19) has the schematic form

$$R_T \leq \mathcal{O}\left(\frac{a}{\sqrt{T}} \right) + \mathcal{O}\left(\frac{1}{T} \right) + \mathcal{O}\left(\frac{b}{qT} \right), \tag{216}$$

where

$$a := \Delta_0 + \frac{d\sigma^2}{N}, \qquad b := \frac{d^2(G^2 + \sigma^2)}{(1-\rho^2)^2}. \tag{217}$$

Substituting (215) into (216) gives

$$R_{\mathcal{B}}^{\mathrm{COSMO}} \leq \mathcal{O}\left(a\sqrt{\frac{qB_{\mathrm{val}}}{\mathcal{B}}} \right) + \mathcal{O}\left(\frac{qB_{\mathrm{val}}}{\mathcal{B}} \right) + \mathcal{O}\left(\frac{d^2(G^2 + \sigma^2)B_{\mathrm{val}}}{(1-\rho^2)^2 \mathcal{B}} \right). \tag{218}$$

The network term is first-order independent of q after converting rounds to bits: larger supports mix more coordinates per round and consume proportionally more communication. The other displayed terms favor the additional rounds purchased by smaller supports. Appendix A.19 separates this prediction from query cost and finite-precision considerations in practical parameter selection.

To isolate representation cost, consider transmitting the same q-sparse round with an explicit coordinate list. With the same value precision, the budget purchases

$$T_{\mathrm{idx}} = \left\lfloor \frac{\mathcal{B}}{q(B_{\mathrm{val}} + \ell_d)} \right\rfloor \approx \frac{T_{\mathrm{COSMO}}}{1 + \ell_d / B_{\mathrm{val}}}. \tag{219}$$

Thus index metadata directly reduces the number of rounds available under the same bit budget. If the update sequence has the same iteration-wise bound and differs only in its encoding, the leading

stochastic term is inflated by

$$\sqrt{1+\frac{\ell_d}{B_{\mathrm{val}}}}, \tag{220}$$

and the lower-order $1/T$ terms are inflated by

$$1+\frac{\ell_d}{B_{\mathrm{val}}}. \tag{221}$$

These factors quantify the effect of encoding the same update sequence. For full-model Qwen with 32-bit values, (221) is 2.03 before any dense refresh is counted. Seed-aware sparse encodings provide a second value-only reference, as evaluated in the Qwen controls.

The same calculation covers a mask refreshed every K rounds. A coordinate-list implementation has amortized multiplier

$$m_{\mathrm{adapt}}(K)=1+\frac{\ell_d}{KB_{\mathrm{val}}}, \tag{222}$$

while any zero-error code is lower bounded by $1+H(S\mid R)/(KqB_{\mathrm{val}})$. For full-model Qwen, the coordinate-list factors are 2.03, 1.10, and 1.02 for $K=1,10,50$. Long mask reuse therefore makes metadata inexpensive when one stable subspace is adequate. ZO-COSMO addresses the complementary case in which supports continue to rotate for coordinate coverage and no data-dependent mask-selection state is maintained.

For the ZO-MARINA-type baseline, combining (208) with the value-only ZO-COSMO cost gives the per-round cost multiplier

$$m_{\mathrm{MARINA}}=\frac{C_{\mathrm{MARINA}}}{qB_{\mathrm{val}}}=\frac{K-1}{K}\left(1+\frac{\ell_d}{B_{\mathrm{val}}}\right)+\frac{d}{Kq}. \tag{223}$$

The first term is the indexed sparse overhead; the second term is the amortized dense refresh. Under a fixed bit budget, ZO-COSMO can take approximately m_{MARINA} times as many sparse rounds as this ZO-MARINA-type baseline.

Table 7: Round-cost multipliers relative to ZO-COSMO.

Setting	d	q	Indexed sparse multiplier	ZO-MARINA-type multiplier, $K=25$
Qwen LoRA-scale	2.52×10^6	8,192	$1.69\times$	$13.94\times$
Qwen full-model	6.53×10^9	8,192	$2.03\times$	$\approx 3.19\times10^4\times$

The refresh cost in (223) grows as $d/(Kq)$. The table uses the LoRA period $K=25$; its full-model row extrapolates payload cost. In Qwen LoRA, the first dense packet exceeds the fixed comparison budget, so the corresponding entry reports initialization. Equal-round SST-2 results separately evaluate the trained refresh method with its full cumulative payload.

D Experimental Details and Reproducibility

We report all experiments using cumulative directed-link communication. For a target budget, each curve is evaluated at the latest logged checkpoint whose cumulative bit count does not exceed that budget. This avoids interpolating across different communication schedules. Sparse baselines that reveal their chosen coordinates are charged for both values and coordinate identifiers. ZO-COSMO is charged only for the transmitted values after the shared seed is fixed. For the ZO-MARINA-type baseline, we charge indexed sparse communication on ordinary rounds and the full dense vector on refresh rounds.

D.1 Fixed-Budget Evaluation Protocol

Every run logs an iteration counter, a method name, a seed, the graph topology, the active support size q, the target metric, and the cumulative communication cost per node. Let r index a single method/seed/topology/q run and let

$$\{(b_{r,0},m_{r,0}),\dots,(b_{r,L_r},m_{r,L_r})\} \tag{224}$$

be its logged cumulative-bit and metric pairs, sorted by $b_{r,\ell}$. For a target budget B, the reported fixed-budget value is

$$\widehat{m}_r(B) = m_{r,\ell^\star}, \qquad \ell^\star = \max\{\ell : \ b_{r,\ell} \le B\}. \tag{225}$$

If no checkpoint satisfies the budget, the run is excluded from that budget rather than extrapolated backwards. In the reported tables this case does not occur for the displayed budgets. The aggregate value is then the mean and standard deviation of $\widehat{m}_r(B)$ over seeds:

$$\bar{m}(B) = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \widehat{m}_r(B), \qquad s(B)^2 = \frac{1}{|\mathcal{R}|-1}\sum_{r\in\mathcal{R}} \big(\widehat{m}_r(B) - \bar{m}(B)\big)^2. \tag{226}$$

The selected checkpoint lies within the budget. We use its measured value directly, without interpolating between different communication schedules.

The per-node communication counter follows the graph. A directed message with payload cost C sent over every directed edge contributes approximately $\bar{d}_{\mathrm{graph}} C$ bits per node, where $\bar{d}_{\mathrm{graph}}$ is the average degree. The code records this value through the same accounting routine for every method. Thus a sparse value-only round, an indexed sparse round, and a dense refresh differ only through the payload formula:

$$C_{\mathrm{value}} = qB_{\mathrm{val}}, \qquad C_{\mathrm{indexed}} = q(B_{\mathrm{val}} + \ell_d), \qquad C_{\mathrm{dense}} = dB_{\mathrm{val}}. \tag{227}$$

Fixed-budget plots use cumulative bits; equal-round controls show the corresponding optimization dynamics.

Algorithm 4 Fixed-Budget Table Extraction

Require: Raw logs with cumulative bits $b_{r,\ell}$ and metric $m_{r,\ell}$; budgets $\mathcal{B}$
1: **for** budget $B \in \mathcal{B}$ **do**
2: **for** each method/seed/topology/q run r **do**
3: Select $\ell^\star = \max\{\ell : b_{r,\ell} \le B\}$
4: Record $\widehat{m}_r(B) = m_{r,\ell^\star}$
5: **end for**
6: Report the seed mean and standard deviation of $\widehat{m}_r(B)$
7: **end for**

Table 8: Core experimental hyperparameters.

Experiment	d	N	q	Steps	η	μ
Rosenbrock main	20	10	1	3000	2.5×10^{-3}	5×10^{-3}
Rosenbrock q sweep	20	10	$\{1,2,4,8,16\}$	3000	8×10^{-4}	5×10^{-3}
Network stress	128	$\{4,8,16,32,64\}$	16	1200	4×10^{-4}	5×10^{-3}
Support locality	128	$\{8,16,32,64\}$	16	2200	4×10^{-4}	5×10^{-3}
Qwen LoRA confirmatory	2.52×10^{6}	2	8192	59/100	5×10^{-6}–5×10^{-5}	10^{-3}
Qwen topology diagnostic	2.52×10^{6}	$\{2,4,8\}$	8192	100	5×10^{-5}	10^{-3}
Qwen edge-local study	2.52×10^{6}	$\{4,8\}$	8192	14–100	5×10^{-5}	10^{-3}
Qwen full model	6.53×10^{9}	1	8192	200	10^{-8}	10^{-3}

Unless otherwise stated, sparse payloads use 32-bit values and momentum uses $\beta = 0.9$. The Rosenbrock main comparison selects η from $\{8\times10^{-4}, 1.2\times10^{-3}, 1.6\times10^{-3}, 2\times10^{-3}, 2.5\times10^{-3}, 3\times10^{-3}\}$ on pilot seed 0 and reports disjoint seeds 1–5. The support-size ablation fixes one learning rate. The network stress and support-locality tests use the core method ($\beta = 0$) and heterogeneity scale 0.3. Edge supports are generated from $(s_0, t, \min\{i,j\}, \max\{i,j\})$; no edge identifier or coordinate list is transmitted.

The confirmatory QNLI study uses $\eta = 5\times10^{-5}$ for both ZO-COSMO and Rand-k. The zero-momentum controls use either that nominal step or 5×10^{-6} to match the initial scale $(1-\beta)\eta$. ZO-MARINA-type uses 2.5×10^{-5}. The topology and edge-local studies use five seeds and charge every repeated exchange. The fixed-budget study uses 26.214M bits per node. The same-matching extension runs to 1540 rounds with either a constant step or a common tenfold reduction after round 100; all other optimization settings are unchanged. Both report full-validation accuracy and label NLL.

The Rosenbrock experiment permits direct runs of both sparse and dense baselines on the same nonconvex objective and query budget. The indexed control isolates recurring metadata cost, while Com-DSZO tests adaptive, error-compensated compression.

For node i, the local Rosenbrock objective is

$$f_i(x) = \sum_{r=1}^{d-1} \left[\alpha \left(z_{i,r+1} - z_{i,r}^2 \right)^2 + (1 - z_{i,r})^2 \right], \qquad z_i = x - s_i, \tag{228}$$

where $\alpha = 2$ and s_i is a small node-specific shift. The shift scale is 0.02 in the reported runs. This creates mild heterogeneity without changing the qualitative geometry of the objective. We report both the objective at the network average and the average local objective in the logs; the main paper uses the objective at the network average because it is the quantity controlled by the convergence theorem.

All Rosenbrock runs use the same initialization, smoothing radius, value precision, graph, and oracle budget within a seed. In the main comparison, the indexed control copies the ZO-COSMO support and update exactly and changes only the message cost. Com-DSZO instead uses dense local ZO directions and Top-k innovation compression as defined in Appendix C.6. The separate support-size ablation retains the earlier independent Rand-k comparator, a no-momentum control, and dense ZO-DSGD.

Table 9: Rosenbrock q-sensitivity at fixed bit budgets with the common step size $\eta = 8 \times 10^{-4}$. Cells report ZO-COSMO / indexed Rand-k, mean $\pm$ standard deviation over five seeds; lower is better.

q	0.3M bits	1M bits	3M bits
1	0.131 ± 0.018 / 0.145 ± 0.017	0.119 ± 0.009 / 0.113 ± 0.006	0.119 ± 0.009 / 0.113 ± 0.006
2	0.632 ± 0.314 / 0.866 ± 0.384	0.121 ± 0.011 / 0.113 ± 0.008	0.121 ± 0.011 / 0.113 ± 0.008
4	7.283 ± 0.322 / 8.054 ± 0.540	0.172 ± 0.023 / 0.192 ± 0.023	0.121 ± 0.009 / 0.115 ± 0.009
8	13.971 ± 0.306 / 14.689 ± 0.569	1.630 ± 0.594 / 2.679 ± 0.947	0.124 ± 0.012 / 0.122 ± 0.015
16	22.680 ± 3.168 / 23.303 ± 0.827	9.385 ± 0.857 / 10.234 ± 0.643	0.288 ± 0.058 / 0.401 ± 0.131

At $q = 1$ and $q = 2$, both methods take many rounds at the larger budgets and approach saturation. At $q = 8$ and $q = 16$, larger packets leave fewer rounds and the fixed-budget gap is clearer (Table 9). The sweep illustrates how the value of compact encoding changes with support size, complementing the identical-update control in the main experiment.

For all decentralized experiments, we use symmetric connected graphs and Metropolis mixing weights. Given an adjacency matrix A, the off-diagonal weights are

$$W_{ij} = \frac{1}{1 + \max\{\deg(i), \deg(j)\}} \quad \text{if } A_{ij} = 1,\ i \neq j, \tag{229}$$

and $W_{ii} = 1 - \sum_{j \neq i} W_{ij}$. The reported spectral parameter is computed from the centered mixing matrix:

$$\rho = \rho_{\mathrm{spec}} \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\top \right), \tag{230}$$

where $\rho_{\mathrm{spec}}(\cdot)$ is the spectral radius. Ring, grid, and Erdős-Rényi graphs therefore differ mainly through this mixing parameter and average degree. The Erdős-Rényi graphs are resampled until connected, with edge probability 0.4.

The network stress test changes both N and the graph family while holding the objective dimension, support size, step size, and horizon fixed. It uses $d = 128$, $q = 16$, $T = 1200$, five seeds, and node shifts of scale 0.3. Complete graphs provide the full-mixing reference; grids grow in two dimensions; rings deliberately drive the spectral gap toward zero. At $N = 64$, their respective ρ values are 0, 0.9677, and 0.9968. All three reduce the objective to about 6.1% of its initial value. The network-average objective remains within 0.1% across graph families, while disagreement records the predicted deterioration on the ring.

D.2 Support-Locality Experiment

This experiment compares the communication operators in Algorithm 2. All methods use the same heterogeneous Rosenbrock network, $d = 128$, $q = 16$, two function queries per node and round, no

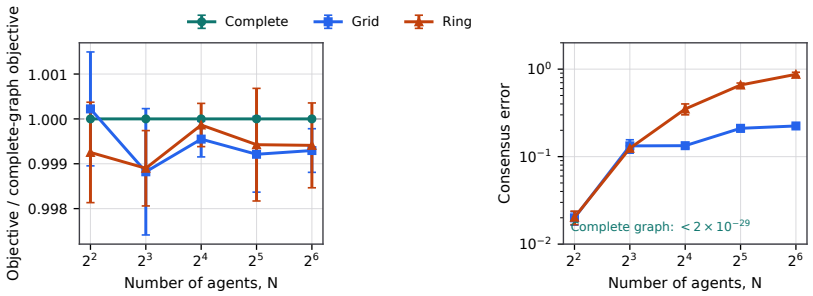


Figure 14: Five-seed topology stress test for the global-support update.

momentum, and five seeds on even rings. Global-support ZO-COSMO sends q values to both ring neighbors, costing $2qB_{\text{val}} = 1024$ bits per node and round. Edge-local ZO-COSMO alternates the ring's two perfect matchings and sends one q-value message, costing 512 bits per node and round.

The seed–scalar control uses ideal all-gather communication. Each node forms a dense seed-reconstructible Rademacher ZO update, and every node applies the exact average after diameter-depth dissemination. We charge the spanning-tree lower bound $(N-1)(64+32)$ bits per node for a 64-bit seed and 32-bit scalar. This operator-level control uses exact aggregation and the minimum packet-edge count from Appendix C.1; SeedFlood's SubCGE implementation is not reproduced.

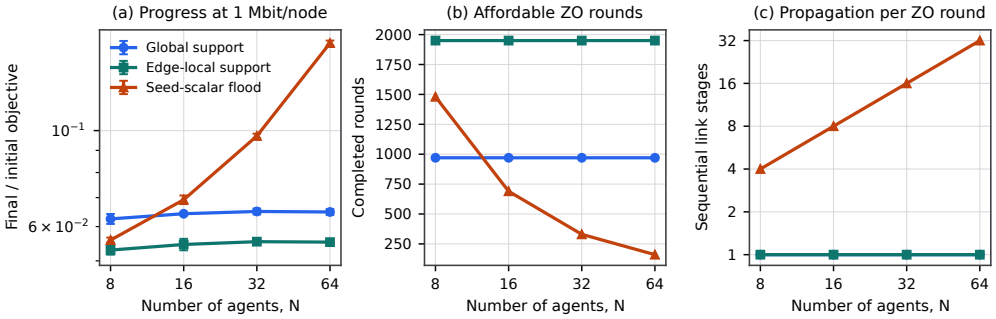


Figure 15: One-hop support locality versus graph-wide seed–scalar dissemination. Bars show one standard deviation over five seeds.

Table 10: Relative objective at 1 Mbit per node. Parentheses give completed logged rounds.

N	Global support	Edge-local support	Seed–scalar flood
8	0.0625 ± 0.0017 (970)	$\mathbf{0.0530 \pm 0.0013}$ (1950)	0.0558 ± 0.0007 (1480)
16	0.0643 ± 0.0003 (970)	$\mathbf{0.0546 \pm 0.0017}$ (1950)	0.0692 ± 0.0016 (690)
32	0.0650 ± 0.0010 (970)	$\mathbf{0.0554 \pm 0.0006}$ (1950)	0.0971 ± 0.0012 (330)
64	0.0649 ± 0.0010 (970)	$\mathbf{0.0552 \pm 0.0004}$ (1950)	0.1593 ± 0.0021 (160)

At $N = 8$, the compact flood packet offsets its four sequential ring stages and remains competitive. Its payload then grows linearly with N and its propagation depth grows as $N/2$: by $N = 64$, only 160 logged rounds fit the budget. Edge-local support uses one message and one neighbor stage per active node and is best at every tested N. Global support is also one-hop, but spends twice the ring payload to mix both neighbors in each round. These results quantify the communication benefit of local mixing as the ring grows.

Same-matching support control. We fix the alternating matching and give every edge the same public support and signs. Both matching variants cost 512 bits per node and round, with identical query counts, initialization, and $\eta = 4 \times 10^{-4}$. Their objectives are close at $N = 8, 16$, separating the scheduling gain from any additional benefit of independent supports. At $N = 4$, the common-support variants are unstable on one of five seeds. Table 11 includes all five seeds with unclipped updates.

Table 11: Same-matching control at 1 Mbit per node, five-seed mean ± std of $F(\bar{x}_T)/F(\bar{x}_0)$.

N	All-neighbor, common	Matching, common	Matching, edge-local
4	$(1.18 \pm 2.63)10^{86}$	$(7.61 \pm 17.01)10^{106}$	0.0466 ± 0.0017
8	0.0626 ± 0.0017	0.0520 ± 0.0013	0.0529 ± 0.0013
16	0.0644 ± 0.0003	0.0543 ± 0.0015	0.0545 ± 0.0017

D.3 QWEN2-7B EXPERIMENTS

Task selection. SST-2 starts near its accuracy ceiling, so we fixed a second-task rule before running any method. The candidate pool was QNLI, BoolQ, MRPC, and RTE: binary accuracy tasks with single-token label verbalizers and at most 6000 validation examples. We evaluated the unmodified Qwen2-7B model on each full validation set, retained tasks with accuracy in $[55, 90)\%$, and selected the eligible task with the largest validation set. Table 12 records the complete screen; the rule selects QNLI.

Table 12: Qwen2-7B task screen before method training.

Task	Validation examples	Initial accuracy	Selected
QNLI	5463	60.79%	Yes
BoolQ	3270	81.01%	No
MRPC	408	69.12%	No
RTE	277	78.34%	No

The confirmatory QNLI study evaluates all 5463 validation examples for ten paired seeds. Its initial accuracy is 60.79%. We fix the directed-link budget at 26.214 Mbit, enough for 100 value-only ZO-COSMO rounds. Indexed Rand-k costs 0.442 Mbit per round and therefore completes 59 rounds without exceeding the same budget. Method selection and optimization settings are paired by seed; the primary endpoint is full-validation accuracy.

Two-worker protocol. The LoRA benchmark simulates two decentralized workers. Each worker has an independent LoRA state and momentum, receives a disjoint contiguous training shard, uses batch size 4 and maximum sequence length 256, and participates in exact complete-graph mixing. The network-average model is evaluated on all 872 SST-2 examples or all 5463 QNLI examples. Rank-8 adapters on the query and value projections give $d = 2{,}523{,}136$ trainable coordinates. We use $q = 8192$, $\mu = 10^{-3}$, and subspace momentum $\beta = 0.9$. ZO-COSMO and Indexed Rand-k use $\eta = 5\times10^{-5}$; the stable $\eta = 2.5\times10^{-5}$ setting is used for ZO-MARINA-type, whose dense refresh period is 25 rounds.

Each worker makes two forward queries on its local minibatch. ZO-COSMO uses one public support and exchanges ordered update values. Rand-k uses independent node supports generated from (s_0, t, i); its original encoding also sends their indices. Both replicas apply the same averaged sparse updates. Common initialization therefore keeps their trainable states identical, even when their momentum and queried supports differ. The $N > 2$ experiments below test disagreement reduction between distinct peer states.

Seed-aware re-encoding. This post-hoc control exposes the existing support-generator seed and node identifier to the receiver. It reconstructs each independent support while preserving the directions, values, update order, and trajectory. An existing 100-round Rand-k run then costs $100qB_{\mathrm{val}} = 26.2144$ Mbit per link instead of $100q(B_{\mathrm{val}} + 22) = 44.2368$ Mbit. The analysis reuses the original training logs. A packet-level test verifies exact float32 trajectory agreement between encodings. Seed setup and packet headers are accounted for separately from recurring value fields.

Table 13: QNLI at $B = 26.214$ Mbit per directed link (ten seeds). Seed-aware Rand-k re-encodes existing runs.

Method	Mbit/round	Rounds	Accuracy (%)
Explicit-index Rand-k	0.442	59	61.75±3.41
Seed-aware Rand-k	0.262	100	66.48±7.36
ZO-COSMO	0.262	100	65.40±3.57

The supplied CSVs contain individual-seed measurements; figures show mean and standard deviation.

Explicit-index Rand-k leaves 0.44% of the budget unused. The paired ZO-COSMO contrast is $+3.65$ points (95% percentile-bootstrap CI $[0.49, 6.86]$, 100,000 resamples). Against seed-aware Rand-k, it is -1.08 points (CI $[-5.85, 3.67]$), with mean query loss lower by 0.091 (CI $[0.006, 0.175]$). The accuracy interval spans zero. ZO-MARINA-type is reported at its 60.79% initialization because its first refresh costs 80.7 Mbit, above the comparison budget.

Table 14: QNLI after 100 rounds (ten seeds). The first two rows are identical trajectories under different encodings.

Method	Rounds	Mbit/link	Accuracy (%)	Query loss
Indexed Rand-k	100	44.237	66.48±7.36	6.700±0.178
Seed-aware Rand-k	100	26.214	66.48±7.36	6.700±0.178
ZO-COSMO	100	**26.214**	65.40±3.57	**6.609±0.190**

Momentum and effective step size. On a coordinate with zero momentum state, the first update is $(1-\beta)g$. Thus removing $\beta = 0.9$ while holding η fixed makes the first update ten times larger. Table 15 reports both that stress test and a control with η reduced by ten, which isolates momentum memory from this deterministic rescaling.

Table 15: QNLI momentum controls after 100 rounds (ten seeds). The matched-step comparison gives a 3.92-point momentum gain, with 95% paired bootstrap CI $[1.67, 6.08]$.

Variant	β	η	$(1-\beta)\eta$	Accuracy (%)
Same nominal step	0	5×10^{-5}	5×10^{-5}	50.40±0.61
Core, matched first step	0	5×10^{-6}	5×10^{-6}	61.48±3.31
ZO-COSMO	0.9	5×10^{-5}	5×10^{-6}	**65.40±3.57**

Table 16: SST-2 after 100 rounds, mean ± standard deviation over five seeds.

Method	η	**Final acc. (%)**	**Last-50 query loss**	**Total Mbit/link**	**Relative**
ZO-COSMO (Ours)	5×10^{-5}	94.13±0.42	0.741±0.165	**26.2**	**1.00×**
Indexed Rand-k	5×10^{-5}	94.15±0.63	**0.725±0.060**	44.2	1.69×
ZO-MARINA-type	2.5×10^{-5}	**94.84±0.24**	0.976±0.301	365.4	13.94×

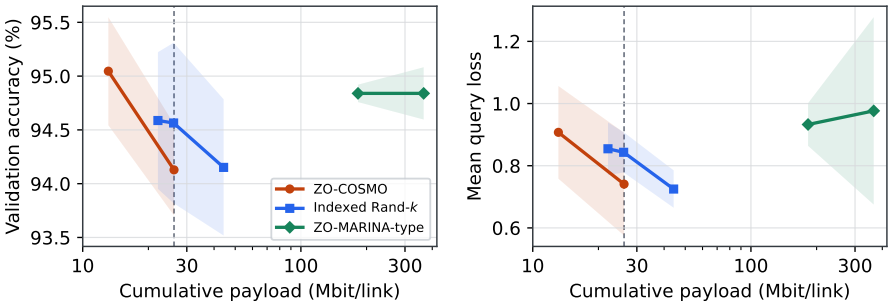
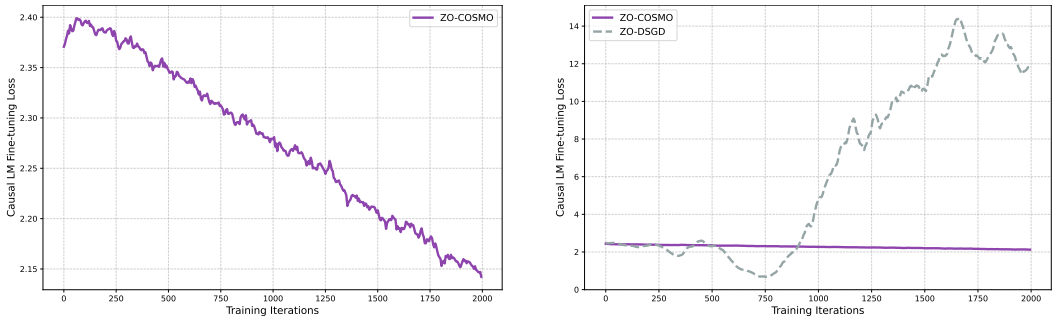


Figure 16: Two-worker Qwen2-7B SST-2 results over five seeds. The dashed line marks the 26.214M-bit comparison budget.



(1) ZO-COSMO. (2) Comparison.

Figure 17: Qwen2-7B LoRA training-loss trajectories on SST-2. ZO-COSMO reaches loss 2.12 rather than 11.46 for dense ZO-DSGD with 308× less total payload.

At the 26.214M-bit budget, ZO-COSMO completes 100 rounds and Indexed Rand-k completes 59 rounds using 26.100M bits. Their accuracies are $94.13 \pm 0.42\%$ and $94.56 \pm 0.75\%$, respectively; both remain in the pretrained model's near-ceiling regime. ZO-COSMO has the lower interval query loss, 0.741 ± 0.165 versus 0.843 ± 0.065. ZO-MARINA-type cannot purchase its first 80.7M-bit dense refresh. The maximum sampled discrepancy between synchronized replicas is zero in every run.

Logical-worker topology simulation. Every logical worker has its own float32 LoRA state, momentum, and disjoint QNLI shard. Logical workers may share an in-memory copy of the immutable bfloat16 Qwen2-7B backbone. A placement audit runs the same four-worker trajectories on one and two backbone copies: all query values, final adapter fingerprints, and validation metrics agree exactly. Repeated runs and an additional intermediate evaluation also reproduce the trajectory exactly; evaluation restores every worker's state. These checks verify the simulated graph updates independently of memory placement.

We run $N \in \{2, 4, 8\}$ logical workers for 100 rounds over complete and ring graphs, using the same $q = 8192$, $\mu = 10^{-3}$, $\beta = 0.9$, batch size 4, and learning rate 5×10^{-5} as the confirmatory run. The reported model is the network-average LoRA state, evaluated on all 5463 QNLI validation examples. One gossip pass costs 26.214 Mbit per directed link. At $N = 8$, this is 1.468 Gbit over the complete graph and 0.419 Gbit over the ring.

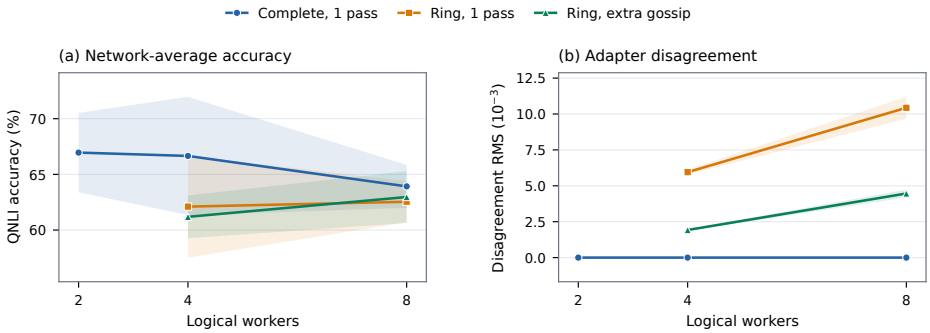


Figure 18: QNLI with 2–8 logical workers after 100 rounds (five-seed mean ± std). Extra ring passes are fully charged.

The complete-graph means are $66.96 \pm 3.57\%$, $66.66 \pm 5.31\%$, and $63.92 \pm 1.93\%$ for $N = 2, 4, 8$. With one ring pass, $N = 4$ reaches $62.10 \pm 4.58\%$ and $N = 8$ reaches $62.54 \pm 1.85\%$. Their consensus RMS values, sampled on a fixed set of 2048 LoRA coordinates, are 0.00595 ± 0.00018 and 0.01043 ± 0.00076. Repeating the exchange twice at $N = 4$ and four times at $N = 8$ changes the spectral factors from 0.333 to 0.111 and from 0.805 to 0.419. RMS falls to 0.00192 ± 0.00005 and 0.00446 ± 0.00028, while accuracy becomes $61.19 \pm 1.93\%$ and $62.97 \pm 2.33\%$. Extra gossip consistently reduces disagreement, but its task-metric benefit is not uniform.

Edge-local QNLI study. We next compare global support with Algorithm 2 on complete and ring overlays for $N \in \{4, 8\}$. The budget is 26.214 Mbit per node. A global round sends one q-value message on every outgoing edge, so the last feasible checkpoints are 33 and 14 rounds on the $N = 4$ and $N = 8$ complete overlays, and 50 rounds on either ring. A perfect-matching round sends one message per node, allowing 100 edge-local rounds. The index-charged control is the same edge-local trajectory at round 59, re-encoded with 22 coordinate bits per value. We report full-validation accuracy and label NLL over five paired seeds; the 100-round control separates communication efficiency from per-round mixing strength.

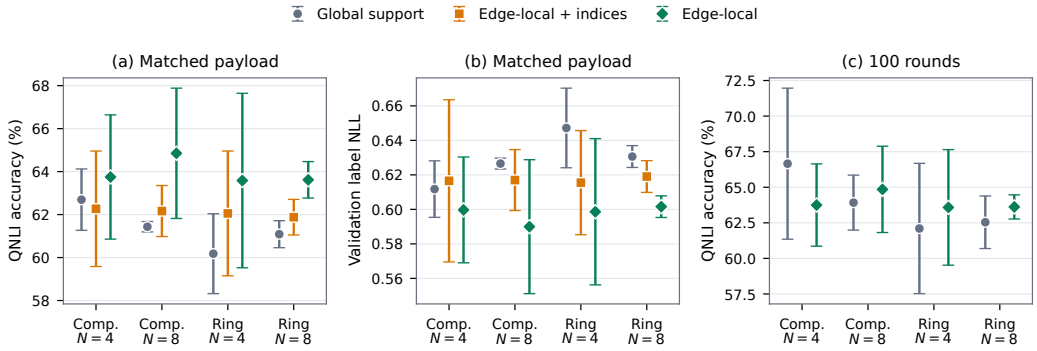


Figure 19: QNLI with global and edge-local support (five-seed mean ± std). The first two panels fix 26.214 Mbit per node; the last fixes 100 rounds.

On the eight-worker complete graph, global support, index-charged edge-local, and index-free edge-local updates reach $61.44 \pm 0.24\%$, $62.16 \pm 1.19\%$, and $64.85 \pm 3.03\%$ accuracy. The index-free gain over global support is 3.42 points (95% paired bootstrap CI $[1.06, 5.63]$), with a label-NLL reduction of 0.0366 (CI $[0.0030, 0.0649]$). On the eight-worker ring, the endpoints are $61.09 \pm 0.63\%$, $61.88 \pm 0.83\%$, and $63.62 \pm 0.85\%$: an accuracy gain of 2.53 points (CI $[1.87, 3.07]$) and an NLL reduction of 0.0291 (CI $[0.0238, 0.0343]$).

At $N = 4$, the corresponding accuracy gains are 1.05 points on the complete graph (CI $[-2.00, 4.10]$) and 3.40 points on the ring (CI $[0.31, 6.49]$). Equal-100-round accuracy intervals span zero in three settings; the eight-worker ring gives $+1.08$ points (CI $[0.07, 2.07]$), while all four equal-round NLL intervals span zero. Matching purchases more updates per bit, whereas all-neighbor mixing reaches more peers on each selected coordinate. The alternating ring satisfies the two-round contraction in (154). These paired intervals describe each comparison separately, without a multiple-comparison adjustment. The next control holds scheduling fixed.

Here $d/q = 308$, so the 100-round Qwen runs have exposure $\tau = qT/d \approx 0.325$. Under iid uniform supports, a coordinate has been selected with probability $1 - (307/308)^{100} \approx 0.278$. The longer control below extends Qwen training to $\tau = 5$. The quadratic mechanism study reaches $\tau = 200$, covering substantially more repeated coordinate activations.

Fixed-matching controls. On the four-worker alternating ring, G shares support and signs across matched pairs, S shares support with independent edge signs, and I uses independent edge supports and signs. G/S retain identical support draws; S/I retain identical ordered edge-sign draws. All arms use the same five seeds, initialization, shards, $q = 8192$, $\mu = 10^{-3}$, and $\beta = 0.9$. Each worker maintains float32 adapters on the frozen bfloat16 backbone. The terminal cost at $T = 1540$ is 403.702 Mbit and 3080 optimization queries per worker, with exposure $qT/d = 5$.

Scheduling at fixed query correlation. The secondary scheduling control compares matching G at round 100 with all-neighbor global-support mixing at round 50, both at 26.214 Mbit per worker. They use the same query-direction generator and initialization. Label NLL falls from 0.6472 ± 0.0231 to 0.5902 ± 0.0474, a reduction of 0.0570 (paired 95% CI $[0.0172, 0.1050]$). Accuracy is $64.58 \pm 5.40\%$ versus $60.18 \pm 1.86\%$ (paired $+4.40$ points, CI $[-0.01, 9.42]$). At equal 100 rounds, both intervals span zero. The same-budget NLL gain therefore persists with query correlation unchanged.

Long-horizon adaptation. All three couplings use $\eta = 5 \times 10^{-5}$ for the first 100 updates and $\eta = 5 \times 10^{-6}$ thereafter, with momentum carried through the switch. The common schedule, five seeds, and terminal accuracy endpoint were fixed before training. We evaluate the full validation set at 100, 308, 912, 924, and 1540 rounds and reuse the archived constant-step study as the paired reference. All 15 new runs complete. Their round-100 accuracy, NLL, query statistics, disagreement, coverage, and payload reproduce the reference records exactly.

At 1540 rounds, G, S, and I reach $74.07 \pm 2.58\%$, $77.82 \pm 2.80\%$, and $75.97 \pm 2.75\%$ accuracy, respectively. Their gains over round 100 are 9.49, 12.56, and 12.38 points, with paired intervals $[6.79, 11.84]$, $[7.56, 17.08]$, and $[7.87, 16.89]$. Label NLL falls to 0.528 ± 0.022, 0.501 ± 0.031, and 0.528 ± 0.028. Mean disagreement RMS is 0.00826, 0.00804, and 0.00728, while visited-coordinate coverage reaches 99.33–99.34%. Training thus remains effective after repeated coordinate activations, with lower disagreement than at round 100.

The archived constant-step counterparts end at $50.12 \pm 0.69\%$, $49.72 \pm 0.23\%$, and $50.00 \pm 0.96\%$, with disagreement RMS 1.65–1.71. The shared step reduction improves terminal accuracy by 23.95, 28.10, and 25.96 points. Figure 20 shows all scheduled checkpoints of both studies.

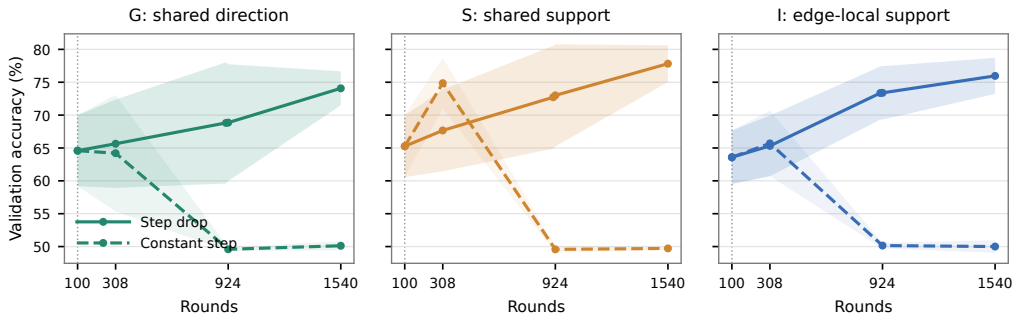


Figure 20: Long-horizon QNLI, five-seed mean ± std. Solid: tenfold step reduction after round 100; dashed: constant step. Matching and random streams are paired.

Direction correlation at fixed support. With matching, support draws, query counts, and payload fixed, independent edge signs raise S above G by 3.75 accuracy points (95% paired CI $[0.24, 6.79]$) and lower label NLL by 0.0272 (CI $[0.0090, 0.0485]$). The gain comes from cross-edge sign correlation under the same training schedule. Moving from S to I lowers disagreement RMS by 0.00077 (CI $[0.00022, 0.00131]$), while the accuracy difference is -1.86 points (CI $[-5.39, 2.21]$). Thus the additional independent supports reduce disagreement without an established accuracy gain in this comparison. Intervals are descriptive paired estimates over five seeds, without multiple-comparison adjustment.

Long-horizon encoding comparison. At 403.702 Mbit per worker, value-only messages afford 1540 rounds, whereas 32-bit values with 22-bit coordinate indices afford 912. Re-encoding the same edge-local trajectory gives $73.36 \pm 4.04\%$ accuracy at the indexed endpoint and $75.97 \pm 2.75\%$ at the value-only endpoint: $+2.61$ points (paired CI $[0.23, 6.32]$). For G and S the corresponding differences are $+5.26$ (CI $[-0.05, 10.57]$) and $+5.11$ points (CI $[-1.27, 12.66]$). All three NLL intervals span zero. The optimizer trajectory is fixed; removing coordinate lists buys the additional rounds.

The single-worker full-model study tests sparse-update stability at 6.53B trainable coordinates and compares two encodings of the same trajectory. With $q = 8192$, each update costs 262.1K bits with public support or 532.5K bits with explicit coordinates.

In the Qwen full-model study, each step samples a public sparse Rademacher direction supported on q coordinates of the selected trainable parameter set. The two-query loss difference is

$$\widehat{\partial}_t = \frac{\ell(\theta_t + \mu u_t; \xi_t) - \ell(\theta_t - \mu u_t; \xi_t)}{2\mu}, \tag{231}$$

and the communicated sparse values correspond to

$$\widehat{g}_{t,S_t} = \frac{d}{q}\widehat{\partial}_t u_{t,S_t}. \tag{232}$$

Momentum is stored only on coordinates that have appeared in the sparse support, so the auxiliary optimizer state scales with the visited support rather than with the full parameter dimension. In the full-model run, the model itself is loaded in bfloat16, while the sparse momentum map can remain on CPU and grow with the number of distinct sampled coordinates.

The full-model study uses the local Qwen2-7B checkpoint, SST-2 batches of size 8, maximum sequence length 256, and evaluation on the SST-2 validation set. The trainable coordinate set contains all model parameters except token embeddings and the language-model head, giving 6.53×10^9 coordinates. For $q = 8192$, the value-only payload is

$$qB_{\text{val}} = 8192 \cdot 32 = 262{,}144 \quad \text{bits per step}, \tag{233}$$

whereas an indexed sparse message additionally pays

$$q\lceil \log_2(6.53\times10^9)\rceil = 8192 \cdot 33 = 270{,}336 \quad \text{index bits per step}. \tag{234}$$

Adding coordinate identifiers to the same updates gives 532.5K bits per step, $2.03\times$ the value-only payload. Supports, signs, and values are unchanged, so the difference is the support-description cost.

This single-worker full-model run verifies sparse two-query updates at the advertised dimensionality, where coordinate metadata is already comparable to the value payload. The multi-worker LoRA experiments above evaluate decentralized optimization and graph topology.

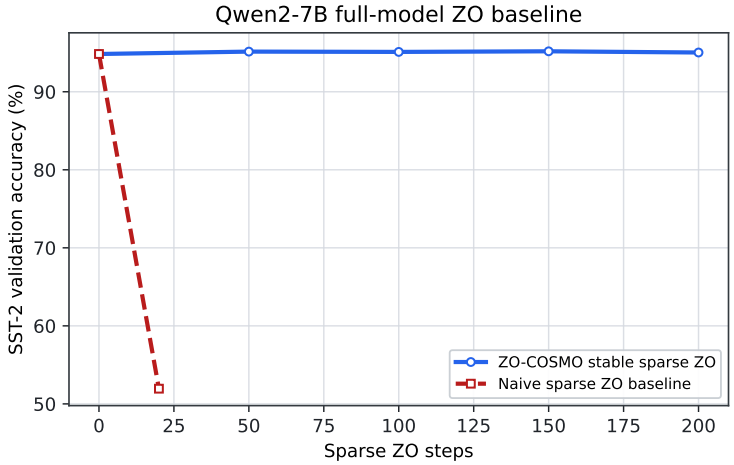


Figure 21: Qwen2-7B full-model sparse ZO study on SST-2.

Table 17: Qwen2-7B full-model sparse ZO communication check.

Run / accounting	Steps	Outcome	Bits / Step	Relative Payload
ZO-COSMO (Ours)	200	**94.84% → 95.03±0.13%**	**262.1K**	**1.00×**
Same ZO-COSMO trajectory + explicit indices	200	identical after re-encoding	532.5K	2.03×
Naive sparse ZO, large step	20	94.84% → 51.95%	262.1K	1.00×

All random-support methods use 32-bit payload accounting. The index fraction $\ell_d/(B_{\text{val}} + \ell_d)$ increases at smaller value widths (Table 2); those lower-bit entries are payload calculations, not quantized training results.

D.4 Reproducibility Artifacts

Raw logs record the method, seed, topology when applicable, support size, metrics, and cumulative communication cost. Tables use the fixed-budget rule in Appendix D.1; figures are generated from the same logs.

Table 18: Main artifacts used for the reported experiments.

Experiment	Config	Raw or summary logs	Main figure/table artifact
Rosenbrock	main, strong-baseline, network-stress, and support-locality YAMLs	fixed-budget, N-sweep, and support-locality raw logs	composite synthetic figures and summary CSVs
Correlation mechanism	original and extension YAMLs	protocols, trajectories, and frozen moments	eight-seed moment and quadratic plots; plot audits
Qwen2-7B	two-worker, topology, edge-local, and full-model YAMLs	ten-seed QNLI, edge-local, SST-2, and full-model logs	LoRA communication, topology, and full-model figures
Communication	serialization and transfer scripts	serialization and 31-trial transfer CSVs	appendix communication diagnostics

The original mechanism study uses `configs/correlation_v21.yaml`, with raw CSVs and its source manifest in `results/correlation_v21/`. The extension uses `configs/correlation_v23.yaml` and `results/correlation_v23/`; its main panels were fixed before training, and the additional appendix views have a separate plot audit. Paths are relative to the experiment-code directory.

The uniform-precision Qwen topology and edge-local study uses `configs/qwen_precision_recheck_v25.yaml`; all 55 trials and 95 checkpoints are retained in `results/qwen_precision_recheck_v25/`. Aggregation checks initialization fingerprints, source hashes, checkpoint completeness, and payload formulas. The longer fixed-matching study uses `configs/qwen_qnli_correlation_v25_fp32.yaml` and `results/qwen_correlation_v25_fp32/`. Each trial records its query trace, coordinate coverage, evaluations, and final adapter states. The evaluation and placement audits are stored separately.

The common step-drop extension uses `configs/qwen_qnli_step_drop_v27.yaml`. All 15 runs and 75 checkpoints are in `results/qwen_step_drop_v27/`; `scripts/summarize_qwen_step_drop.py` verifies the archived reference hashes, exact round-100 matches, recorded learning rates, and payload counters before producing the figures and paired statistics. All three couplings use the same schedule.

The support generator folds the seed with the round number and, when needed, a node or unordered edge identifier. Global-support ZO-COSMO uses no node identifier; edge-local ZO-COSMO agrees within each matched pair. Independent Rand-k uses node identifiers. Those supports are private only when the generator seed is withheld: making it public permits the seed-aware encoding without changing the independent-support trajectories.

The Qwen configuration specifies the HuggingFace checkpoint and trainable-coordinate filter. The topology runner records logical workers, graph, spectral factor, active links, per-node and total payload, and sampled consensus error. The full-model run excludes embeddings and the language-model head, uses bfloat16 weights, and stores sparse momentum in a CPU dictionary. These choices reduce memory use while preserving the recorded protocol and payload.

D.5 Implementation Details

Global support is keyed by the master seed and round; edge-local support additionally uses the unordered endpoints of the active edge. The same support serves queries and message decoding. The explicit-index baselines send coordinate lists, while the seed-aware control reconstructs independent supports from node-specific public keys.

For dense baselines, communication is charged as a full vector payload even if a local implementation stores tensors in contiguous blocks. For sparse baselines, communication is charged per directed

link and per sparse vector. If an undirected graph edge carries messages in both directions, the total edge traffic is twice the directed-link number. The main method comparisons use directed-link cost. The edge-local topology study instead fixes payload per node because global and matching rounds activate different numbers of links; total network payload is recorded as well.

Fine-tuning payloads are counted with 32-bit values. Table 2 gives the corresponding representation costs at other value widths.

Qwen LoRA-scale experiment. Dense ZO-DSGD communicates all LoRA parameters; ZO-COSMO sends only public sparse values. Both use forward-only loss differences. Each logical worker keeps independent float32 adapter and momentum states; the frozen backbone uses bfloat16. Adapter precision is set explicitly when each adapter is created and reset. Sharing a frozen backbone changes memory placement and scheduling, not these states or the graph update.

Qwen full-model experiment. The trainable set excludes token embeddings and the language-model head. Weights use bfloat16, and a coordinate map stores sparse momentum without a dense update buffer. The stable runs use seeds $0, 1, 2$ and are re-encoded with explicit indices; the oversized-step row is a one-seed failure check. Table 3 gives the cost of a hypothetical dense refresh.

D.6 Ablation Design

The momentum ablation keeps the public support and removes the momentum buffer; its matched-first-step control accounts for the resulting scale change. The independent-support ablation changes cross-node alignment. We evaluate both its original coordinate-list encoding and the seed-aware encoding reconstructed from node-specific public seeds.

The same-update indexed control isolates representation: it uses the identical public support, estimator, momentum, and masked consensus as ZO-COSMO, then charges the coordinate list that an ordinary sparse format would carry. Com-DSZO tests a different question. Its Top-k innovation is data-adaptive and error compensated, but it communicates indices and keeps two dense auxiliary arrays in addition to its compressed innovation. Comparing these methods separates an index-free representation gain from the optimization and memory benefits of a stronger compressor.

The periodic-refresh control's first dense packet exceeds the Qwen LoRA comparison budget, so that entry is marked at initialization. Its trained equal-round SST-2 result is reported separately with the full cumulative payload.

E Broader Impact and Ethics

Forward-only fine-tuning reduces storage of backward-pass activations. Local data and exchanged model states still require access controls; function-value access and decentralized execution alone provide no privacy guarantee. The experiments evaluate optimization and payload use with logical workers, while hardware and execution details are reported in Appendix D.