Near-Optimal Decentralized Stochastic Convex Optimization over Networks

Nitai Kluger[*†] **Amit Attia**[*†] **Tomer Koren**[‡]

Abstract

We study decentralized stochastic smooth convex optimization, where M workers (machines) minimize an average objective using local stochastic gradients and neighbor-only communication over a fixed gossip network. A central question in this setting is to determine the largest number of workers that can be used under a total budget of N gradient samples while still preserving the centralized $O(1/\sqrt{N})$ statistical rate. We introduce an accelerated decentralized method that preserves this rate for up to $M \lesssim \sqrt{\rho}\, N^{3/4}$ workers, where ρ is the spectral gap of the gossip network, improving the best prior maximal scaling of $M \lesssim \rho\sqrt{N}$. The method is based on a one-step-delayed stochastic acceleration scheme that enables workers to interleave minibatching with accelerated gossip while controlling residual disagreement, and its guarantee depends only logarithmically on the optimum-local heterogeneity. We establish a matching lower bound for linear-span decentralized first-order methods, showing that the method is optimal up to logarithmic factors.

1 Introduction

Distributed optimization has become a central paradigm in modern large-scale learning and inference. In many applications, data and computation are spread across machines, so optimization must combine many parallel gradient samples with limited communication. The central promise is statistical: for a fixed budget of N gradient samples, using M machines should reduce the number of sequential rounds without degrading the accuracy attainable from those samples.

A useful way to make this promise precise is the *critical parallelism* threshold: the maximum number of machines M for which parallelism preserves the centralized statistical rate. Beyond this threshold, the remaining optimization or communication error is no longer hidden by sampling noise, and additional machines yield diminishing returns for sample efficiency. For smooth convex stochastic optimization, centralized synchronized minibatch SGD preserves the statistical rate for $M \lesssim \sqrt{N}$ workers, while centralized accelerated minibatching raises this scale to $M \lesssim N^{3/4}$ [Dekel et al., 2012, Lan, 2012]. This $N^{3/4}$-scale is therefore the natural benchmark for optimization parallelism.

In this paper we study the same question for decentralized stochastic optimization over networks. In contrast to parameter-server or all-reduce models, decentralized methods communicate only along the edges of a network, formalized as an undirected graph over distributed nodes, typically through gossip averaging, i.e., repeated weighted averaging of vectors between neighbors. This removes a central bottleneck and is a natural abstraction for federated, peer-to-peer, and geographically distributed systems. The price is that workers no longer hold identical iterates: gradients are sampled at different local points, and the disagreement between these points appears as a bias in the stochastic oracle seen by the network average.

[*]Equal contribution.
[†]Blavatnik School of Computer Science and AI, Tel Aviv University; {nitaikluger,amitattia}@mail.tau.ac.il
[‡]Blavatnik School of Computer Science and AI, Tel Aviv University, and Google Research Tel Aviv; tkoren@tauex.tau.ac.il

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

This bias is especially delicate under data heterogeneity. When different workers see different local objectives, their local gradients can point in systematically different directions even when evaluated at the same, nearly optimal model. Thus residual network disagreement is not merely an implementation error; it can create a persistent bias that accelerated methods are known to amplify if it is not controlled carefully [Devolder et al., 2014].

Modern distributed training systems often allow stochastic-gradient computation and neighbor communication to proceed concurrently on the same worker. This motivates the nonblocking scheduling model studied here, where each wall-clock round contains one gradient computation alongside a small, constant number of local gossip steps. Blocking schemes, which spend many consecutive rounds solely on gossiping, may improve agreement, but they idle the stochastic oracles and waste scarce sequential rounds without collecting gradient samples. The relevant challenge is therefore not only to communicate efficiently, but to obtain enough network mixing while continuous parallel sampling is underway.

Several prior works have tackled decentralized stochastic convex optimization using gossip-based stochastic gradients, gradient tracking, and anytime averaging [Koloskova et al., 2020, 2021, Eisen et al., 2025]. To date, the best critical-parallelism guarantee is due to Eisen et al. [2025]: their method preserves the centralized statistical rate for $M \lesssim \rho\sqrt{N}$ workers, where ρ is the spectral gap of the network graph. Thus, on well-connected networks ($\rho = \Theta(1)$), it matches the $\sqrt{N}$-scale of centralized synchronized minibatch SGD, while on general networks it loses a factor depending on the spectral gap.

This leaves a considerable gap to the accelerated $N^{3/4}$-scale. Under a fixed sample budget, the algorithm has only N/M distributed rounds, so the deterministic optimization and network errors must decay fast enough to remain below the $O(1/\sqrt{N})$ statistical term. Non-accelerated stochastic approximation does not provide this decay, and the spectral-gap dependence in existing stochastic decentralized bounds is likewise non-accelerated relative to the optimal deterministic network rates of Scaman et al. [2017]. A separate obstacle is data heterogeneity: the method of Eisen et al. [2025] suffers from a polynomial dependence on heterogeneity, which can dominate the rate. Gradient-tracking methods can remove explicit heterogeneity terms in some regimes [Nedić et al., 2017, Koloskova et al., 2021], but the known stochastic convex bounds imply worse critical parallelism in the sample budget N.

In this paper, we overcome these limitations with a novel delayed accelerated stochastic method that combines two acceleration mechanisms. The optimization dynamics use a Nesterov-type accelerated stochastic recursion, while the network uses accelerated gossip to mix local copies. A one-step-delayed, momentum-compensated query schedule ties these mechanisms together: workers keep sampling gradients at an already available query point while accelerated gossip prepares the next query point from the current local copies. Hence, stochastic acceleration and network mixing are interleaved rather than separated into blocking phases, and the remaining disagreement enters the averaged dynamics as a controlled oracle bias. The method thus performs a single gossip communication step in each round, in addition to one local stochastic gradient query at each node.

The resulting rate is optimal up to logarithmic factors within the standard class of deterministic linear-span (decentralized) first-order methods, with stochastic gradients as the only source of algorithmic randomness. It preserves the optimal stochastic term, attains the optimal network dependence, and depends only logarithmically on the data heterogeneity at the optimum. This gives, to the best of our knowledge, the first accelerated decentralized method in this smooth convex stochastic setting that is rate-optimal up to logarithmic factors.

1.1 Summary of contributions

Our main contributions are summarized as follows:

(i) We propose DDA-SGD, an accelerated method for decentralized stochastic smooth convex optimization over a fixed gossip network, and prove an optimal finite-sample rate for it, up to logarithmic factors. For a total budget of N stochastic-gradient computations over a network with spectral gap $\rho \in (0, 1]$, the output $\widetilde{x}_{T,i}$ of each individual node attains convergence rate

$$\mathbb{E}[f(\widetilde{x}_{T,i}) - f(x^\star)] \lesssim \frac{\sigma R}{\sqrt{N}} + \frac{LR^2M^2}{\rho N^2} \cdot \mathrm{polylog}\left(\frac{N}{\rho}, \frac{\zeta_\star + \sigma}{LR}\right),$$

Method	Iteration complexity to ϵ-accuracy	Maximal parallelism
D-SGD [Koloskova et al., 2020]	$\frac{\sigma^2}{M\epsilon^2} + \frac{\sigma}{\sqrt{\rho}\,\epsilon^{3/2}} + \frac{\zeta_\star}{\rho\,\epsilon^{3/2}} + \frac{1}{\rho\epsilon}$	$\Theta(\sqrt{\rho}\,N^{1/4})$ if $\zeta_\star \approx 0$; $\Theta(\rho N^{1/4})$ otherwise
Gradient Tracking [Koloskova et al., 2021]	$\frac{\sigma^2}{M\epsilon^2} + \frac{\sigma}{c\sqrt{\rho}\,\epsilon^{3/2}} + \frac{1}{c\rho\epsilon}$	$\widetilde{\Theta}(\sqrt{\rho}\,N^{1/4})$ (treating $c = \Theta(1)$)
DAT-SGD [Eisen et al., 2025]	$\frac{\sigma^2}{M\epsilon^2} + \frac{\sqrt{\sigma}+\sqrt{\zeta}}{\rho\epsilon} + \frac{1}{\epsilon}$	$\Theta(\rho\sqrt{N})$
DDA-SGD **This work** (Theorem 1)	$\frac{\sigma^2}{M\epsilon^2} + \frac{1}{\sqrt{\rho\epsilon}} \log \frac{N\zeta_\star}{\rho}$	$\widetilde{\Theta}(\sqrt{\rho}\,N^{3/4})$
Lower bound **This work** (Theorem 2)	$\frac{\sigma^2}{M\epsilon^2} + \frac{1}{\sqrt{\rho\epsilon}}$	No method can exceed $\Theta(\sqrt{\rho}\,N^{3/4})$

Table 1: Comparison of rates for convex decentralized stochastic optimization. Here N is the total number of stochastic gradient samples, M is the number of distributed nodes, ρ is the spectral gap, σ denotes the local noise scale, $\zeta_\star$ denotes data heterogeneity measured at the optimum, ζ denotes a global heterogeneity bound, and c is a negative-eigenvalue parameter appearing in the gradient-tracking analysis of Koloskova et al. [2021]. The table suppresses numerical constants, smoothness/radius parameters, and most logarithmic factors.

where R is an upper bound on the initial distance to the solution, σ^2 bounds the local variance of the stochastic gradients, and $\zeta_\star$ is the data heterogeneity at the optimum. The first term is the centralized optimal stochastic rate, the second term is a network-limited accelerated rate, and $\zeta_\star$ enters only logarithmically. Consequently, the centralized statistical rate is preserved for up to $M \lesssim \sqrt{\rho}\,N^{3/4}$ machines, improving the state-of-the-art $M \lesssim \rho\sqrt{N}$ maximal scaling of the best previous bounds [Eisen et al., 2025]. This improves both the dependence on the global number of gradient computations N and the dependence on the spectral gap ρ.

(ii) We prove a matching lower bound for the network-dependent term. On path networks, any method in the above first-order model needs error at least $\Omega(LR^2M^2/(\rho N^2))$ in the deterministic regime, after translating the communication-round lower bound into the sample-budget notation of our distributed model. Together with the classical stochastic lower bound $\Omega(\sigma R/\sqrt{N})$, this shows that both terms in our upper bound are unavoidable up to logarithmic factors.

(iii) We additionally provide a guarantee for the decentralized stochastic smooth and μ-strongly convex setting via a restarted variant of our method. This restart scheme reaches accuracy ϵ with total sample complexity $N = \widetilde{O}(\sigma^2/\mu\epsilon + M\sqrt{L/\mu\rho}\log(1/\epsilon))$. We defer this extension and its proof to Appendix E.

(iv) A key component of our analysis is a convergence bound for accelerated stochastic optimization with one-step-delayed gradients. The bound shows that, by using momentum-compensated query points, one can retain the usual stochastic accelerated rate even though the gradient needed to form the next standard Nesterov query is not yet available. This delayed-acceleration result may be of independent interest, and is described in Appendix B.

(v) We complement the theory with synthetic least-squares experiments on ring and torus networks, illustrating the optimization performance of DDA-SGD and its scaling with the number of workers relative to D-SGD and DAT-SGD (Section 5).

1.2 Related work

Decentralized stochastic optimization with gossip. Combining gossip averaging with local gradient computation for distributed optimization was pioneered by Nedic and Ozdaglar [2009], who proposed and analyzed distributed subgradient methods for multi-agent convex optimization. Lian et al. [2017] introduced Decentralized Parallel SGD (D-PSGD), showing that decentralized training can match the linear speedup of centralized methods while avoiding the communication bottleneck of parameter servers. Assran et al. [2019] extended gossip-based training to directed and time-varying topologies via stochastic gradient push, improving robustness to stragglers and asymmetric links.

Unified convergence analyses for decentralized SGD under changing topologies were subsequently developed by Koloskova et al. [2020], with rates depending on the spectral gap and data heterogeneity. The gradient-tracking approach of Nedić et al. [2017] and its refined analysis by Koloskova et al. [2021] can eliminate explicit heterogeneity terms at the cost of a weaker dependence on the spectral gap. Variance-reduction and compression techniques have also been studied [Tang et al., 2018a,b, Li et al., 2019]. Most recently, Eisen et al. [2025] combined anytime averaging with accelerated stepsizes to achieve a critical parallelism of $M \lesssim \rho\sqrt{N}$, the best prior scaling. These methods, however, rely on non-accelerated stochastic approximation and therefore fall short of the $N^{3/4}$-scale critical parallelism attainable in centralized accelerated settings.

Classical and centralized acceleration. Acceleration was first developed for centralized first-order optimization, beginning with Nesterov's accelerated gradient method [Nesterov, 1983], which attains the optimal deterministic rate $O(1/T^2)$ for smooth convex optimization after T steps. Lan [2012] extended acceleration to stochastic composite optimization, combining this fast deterministic term with the optimal $O(\sigma/\sqrt{T})$ stochastic rate. Using acceleration in centralized minibatching improves the number of workers that preserve the centralized statistical rate from $M \lesssim \sqrt{N}$ for synchronized minibatch SGD to $M \lesssim N^{3/4}$ [Dekel et al., 2012]. The centralized inexact-oracle analysis of Devolder et al. [2014] shows, however, that accelerated methods are sensitive to persistent oracle errors.

Decentralized and asynchronous acceleration. Accelerated methods have also been developed for deterministic decentralized optimization over networks, including optimal strongly convex rates on fixed networks [Scaman et al., 2017], smooth convex rates [Xu et al., 2020], and time-varying-network guarantees [Rogozin et al., 2020, 2021, Kovalev et al., 2021]. In this line of work, accelerated gossip is incorporated through blocking communication phases [e.g., Scaman et al., 2017, Ye et al., 2023]. Related asynchronous stochastic methods make acceleration compatible with delayed gradients through minibatching, at the cost of filtering out a constant fraction of the samples [Tyurin and Richtárik, 2023, Attia et al., 2025]. Our method instead uses one-step delayed acceleration to exploit every micro-round for both sampling and communication, while controlling the disagreement-induced bias.

Due to space constraints, additional related work is discussed in Appendix A.

2 Problem setup

We first specify the optimization problem, stochastic oracle model, and communication model used by the distributed method. We want to minimize the finite-sum objective

$$f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x),$$

using M decentralized workers. We assume that the solution set is nonempty, and fix a minimizer $x^\star \in \arg\min_x f(x)$. Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^d$.

Worker i has access to a local stochastic gradient oracle $G_i(x, \xi)$ satisfying

$$\mathbb{E}[G_i(x,\xi) \mid x] = \nabla f_i(x), \qquad \mathbb{E}\left[\|G_i(x,\xi) - \nabla f_i(x)\|^2 \mid x\right] \le \sigma^2.$$

These are standard conditional unbiasedness and bounded-variance assumptions. All oracle samples are conditionally independent given past queries.

Bounded data heterogeneity. We assume the local objectives have *bounded gradient heterogeneity at the optimum*: there is a constant $\zeta_\star \ge 0$ such that

$$\frac{1}{M}\sum_{i=1}^{M} \left\|\nabla f_i(x^\star) - \nabla f(x^\star)\right\|^2 \le \zeta_\star^2. \tag{1}$$

This is an optimum-local heterogeneity condition, similar to the assumption used by Koloskova et al. [2020], and should be distinguished from the stronger global bounded-heterogeneity assumptions often used in decentralized stochastic optimization, which require this average squared gradient

mismatch to be uniformly bounded for all x [e.g., Lian et al., 2017, Tang et al., 2018a,b, Li et al., 2019, He et al., 2022, Eisen et al., 2025]. Other relaxations bound heterogeneity at the origin [Lu and De Sa, 2021]. Gradient-tracking methods can completely remove explicit heterogeneity terms in some regimes [Nedić et al., 2017, Koloskova et al., 2021], at the expense of weaker dependence on the spectral gap ρ. Our analysis here does require a finite optimum-local heterogeneity parameter, but $\zeta_\star$ enters the final guarantee only logarithmically.

Decentralized communication. The M workers form a fixed network with an undirected communication graph $\mathcal{G} = (\{1, \dots, M\}, E)$. The basic time unit is a synchronous micro-round. In one micro-round, each worker may perform local computation, make at most one stochastic-gradient oracle call, and perform a small constant number (ideally one) of neighbor communication step; there is no parameter server or all-reduce operation. We restrict attention to nonblocking schedules: algorithms may not insert additional phases consisting of many back-to-back gossip steps with no local oracle work. Thus a budget of N stochastic-gradient computations over M workers corresponds to at most N/M micro-rounds, and communication must be interleaved with those rounds rather than appended in extra blocking phases.

As in the gossip-based decentralized model of Scaman et al. [2017], communication is implemented through a *gossip matrix* $P \in [0,1]^{M\times M}$ supported on the graph, so $P_{ij} = 0$ whenever $i \neq j$ and $\{i, j\} \notin E$. One communication round with P acts as follows: when the local vectors are the columns of $Z \in \mathbb{R}^{d\times M}$, all workers simultaneously exchange messages along the support of P and compute the corresponding columns of ZP. We assume that P is symmetric and doubly stochastic. Formally,

$$P = P^\mathsf{T}, \qquad P\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\mathsf{T}P = \mathbf{1}^\mathsf{T}.$$

Let the eigenvalues of P be ordered as $1 = \lambda_1(P) \geq \lambda_2(P) \geq \cdots \geq \lambda_M(P) \geq -1$. We assume that P has spectral gap at least $\rho \in (0, 1]$, in the sense that

$$1 - \lambda_2(P) \geq \rho > 0. \tag{2}$$

We note that this spectral condition is strictly weaker than the one-step consensus-contraction assumption used in previous work [e.g., Koloskova et al., 2020, 2021, Eisen et al., 2025]. For a fixed symmetric gossip matrix, such a contraction assumption effectively requires $\max\{|\lambda_2(P)|, |\lambda_M(P)|\} < 1$, and therefore separates the most negative eigenvalue of P away from -1. In contrast, our analysis only needs the weaker condition in Eq. (2) and allows $\lambda_M(P)$ to be arbitrarily close to -1.[1]

3 Algorithm

Our method combines Nesterov-type stochastic acceleration, which provides the $O(1/T^2)$ optimization term needed for $N^{3/4}$ critical parallelism, with accelerated gossip to reduce network disagreement. The full procedure is given in Algorithm 1, and its convergence guarantee is stated in Theorem 1. We first describe the macro-step structure conceptually, before giving the full pseudocode.

The method runs in macro steps of B synchronous micro-rounds. At the beginning of macro step t, the algorithm has three states: a primal state X_t, a dual/gradient-accumulation state Z_t, and a query state Q_t. Over T macro steps and M workers, the method uses $N = MBT$ stochastic-gradient samples. Conceptually, each macro step pipelines two activities: workers sample at the already prepared delayed query Q_t, while they gossip the current primal state X_t so it can define the next query. The important departure from standard Nesterov-type stochastic acceleration is that Q_t is not formed from the current primal state. It is a one-step-delayed query, computed from previously mixed primal copies. Appendix B analyzes this technique in the classical single-node setting.

This delay lets one macro step use its B micro-rounds for two purposes at once. First, every worker samples B stochastic gradients at the already available query point $q_{t,i}$, forming a local minibatch and using the full sampling budget. Second, during the same B micro-rounds, the workers run accelerated gossip on the current primal copies X_t. The gossip block is Nesterov acceleration applied to the quadratic disagreement objective associated with $I - P$; see Appendix C. Thus the point being sampled is fixed and available, while the point being communicated is prepared for the next query.

[1] The distinction is already visible on the two-node graph with $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Here P is symmetric and doubly stochastic, $\lambda_2(P) = -1$, and therefore $1 - \lambda_2(P) = 2$, so Eq. (2) holds for every $\rho \leq 1$; however, for the row vector $z = (1, -1)$, we have $zJ = 0$ and $zP = (-1, 1)$, so $\|zP - zJ\|_2 = \|z - zJ\|_2$. Thus no strict one-step consensus contraction holds.

By the end of the macro step, gossip has produced a nearly synchronized primal copy $\widetilde{X}_t$, which is then used to form Q_{t+1}. We emphasize that Q_{t+1} does not depend on the current minibatch gradient; the subsequent state update is local and requires no extra communication or synchronization.

The one-step delay is what makes this separation possible: sampling does not wait for communication, and communication does not disturb the point being queried. The price is that the query sequence must still behave like a Nesterov query sequence. For this reason Q_{t+1} includes a momentum compensation term, which anticipates the next inertial displacement and makes the delayed query mismatch controllable by the same stability terms that drive the accelerated analysis.

Algorithm 1 Decentralized Doubly Accelerated SGD (DDA-SGD)

Input: common $x_0 \in \mathbb{R}^d$, macro horizon T, local minibatch/accelerated-gossip length B, symmetric doubly stochastic gossip matrix P, gossip stepsize $\alpha_{\mathrm{g}} \in (0, 1]$, gossip momentum $\beta_{\mathrm{g}} \in [0, 1)$, weights $(a_t)_{t=0}^{T-1}$, averaging parameters $(\theta_t)_{t=0}^{T}$, momentum parameters $(m_t)_{t=0}^{T+1}$, and dual stepsizes $(\eta_t)_{t=0}^{T}$.
Set $X_{-1} = X_0 = Z_0 = Q_0 = \widetilde{X}_{-1} = x_0\mathbf{1}^\top$.
for $t = 0, 1, \ldots, T$ **do**
  Initialize $Y_t^{(-1)} = Y_t^{(0)} = X_t$ and $g_{t,i} = 0$ at each node i.
  for $s = 1, 2, \ldots, B$ **do**
    Each node i samples one gradient at its current delayed query $q_{t,i}$, the ith column of Q_t, and accumulates

$$g_{t,i} \leftarrow g_{t,i} + \frac{1}{B} G_i(q_{t,i}, \xi_{t,i}^{(s)}).$$

    In the same round, perform one Nesterov accelerated-gossip step:[2]

$$V_t^{(s)} = Y_t^{(s-1)} + \beta_{\mathrm{g}}\big(Y_t^{(s-1)} - Y_t^{(s-2)}\big), \qquad Y_t^{(s)} = (1 - \alpha_{\mathrm{g}})V_t^{(s)} + \alpha_{\mathrm{g}} V_t^{(s)} P.$$

  end for
  Set the accelerated mixed primal copy $\widetilde{X}_t = Y_t^{(B)}$, and form the next delayed query matrix

$$Q_{t+1} = \widetilde{X}_t + m_t(1 + m_{t+1})(\widetilde{X}_t - \widetilde{X}_{t-1}),$$

  whose columns are $q_{t+1,i}$.
  Set $G_t = [g_{t,1}, \ldots, g_{t,M}]$ and update local dual and primal copies:

$$Z_{t+1} = Z_t - \eta_t G_t, \qquad X_{t+1} = (1 - \theta_t)X_t + \theta_t Z_{t+1}.$$

end for
Each node i outputs its local column $\widetilde{x}_{T,i}$ of $\widetilde{X}_T$.

Previous decentralized methods [Koloskova et al., 2020, Eisen et al., 2025] tightly couple the variables used for communication with the variables used for gradient computation. Although such coupling can be useful, our one-step-delayed construction deliberately separates the two roles. This separation is what allows us to combine accelerated gossip with accelerated optimization, two tools that are powerful but sensitive to inexactness [Devolder et al., 2014]. An alternative way to accommodate these techniques is to reserve some rounds purely for communication, effectively leaving a constant fraction of the available gradient queries unused, as in the approaches of Tyurin and Richtárik [2023], Attia et al. [2025]. Such blocking communication phases are outside the nonblocking round model above; the one-step-delayed scheme avoids them because every micro-round can be used both for sampling a stochastic gradient and for performing a gossip step.

4 Main results

Next we present our convergence guarantee for Algorithm 1, and the matching lower bound for linear-span decentralized methods.

[2] Only multiplication by P requires neighbor communication. Scalar multiplications and additions/subtractions are local, so each recurrence step uses one gossip communication round.

Theorem 1 (Distributed delayed-gossip rate)**.** *Let each f_i, $i = 1, \dots, M$, be convex and L-smooth, with an unbiased oracle G_i of conditional variance at most σ^2. Assume conditionally independent oracle samples and optimum-local heterogeneity bounded by $\zeta_\star$, and let P be a fixed symmetric doubly stochastic gossip matrix with spectral gap at least $\rho \in (0, 1]$. Let $f := \frac{1}{M}\sum_{i=1}^{M} f_i$, $x^\star \in \arg\min_x f(x)$, and $R > 0$ satisfy $R \geq \|x_0 - x^\star\|$. Fix a total sample budget N, and define*

$$\Lambda_N := \frac{100N}{\sqrt{M}}\left(1 + \frac{\zeta_\star + \sigma}{LR}\right).$$

Choose $B := \left\lceil \max\{1,\ (20/\sqrt{\rho})\log(\Lambda_N/\rho)\} \right\rceil$, and assume for simplicity that $T := N/(MB)$ is a positive integer. Set the accelerated gossip parameters $\alpha_{\mathrm{g}} := \frac{1}{2}$, $\beta_{\mathrm{g}} := (1 - \sqrt{\rho/2})/(1 + \sqrt{\rho/2})$, and the algorithmic parameters

$$a_t := \frac{t+1}{128L}, \quad A_t := \sum_{i=0}^{t} a_i = \frac{(t+1)(t+2)}{256L}, \quad \theta_t := \frac{a_t}{A_t} = \frac{2}{t+2} \ \ (0 \leq t < T), \quad \theta_T := 0,$$

$$m_0 := 0, \quad m_t := \frac{t-1}{t+2} \ \ (1 \leq t \leq T), \quad m_{T+1} := 0,$$

$$H_{T-1} := \sqrt{\textstyle\sum_{i=0}^{T-1} a_i^2}, \quad \beta := 1 + \frac{\sigma}{R\sqrt{MB}} H_{T-1}, \quad \eta_t := \frac{a_t}{\beta} \ \ (0 \leq t < T), \quad \eta_T := 0.$$

Run Algorithm 1 with these parameters for T macro steps, and let $\widetilde{x}_{T,i}$ be the output held by node i. Then, for every node i,

$$\mathbb{E}[f(\widetilde{x}_{T,i}) - f(x^\star)] \leq \frac{14\sigma R}{\sqrt{N}} + \frac{2^{17} LR^2 M^2}{\rho N^2} \log^2 \frac{\Lambda_N}{\rho}.$$

The optimization parameters adapt a standard Nesterov stochastic-acceleration schedule to the one-step-delayed, momentum-compensated query. The gossip parameters $\alpha_{\mathrm{g}}, \beta_{\mathrm{g}}$ are standard Nesterov choices for the consensus quadratic. The block length B makes residual disagreement small and also acts as a minibatch size, bounding the averaged conditional oracle variance by $\sigma^2/(MB)$.

We give a proof sketch in Section 4.1 and defer the full proof of Theorem 1 to Appendix D. Next, we state a lower bound for linear-span methods, showing the guarantee is nearly optimal.

Theorem 2 (Lower bound for linear-span methods)**.** *There is a universal constant $c > 0$ such that the following holds. For every $M \geq 6$, $L > 0$, $R > 0$, and $S \geq 1$, on the M-node path graph with spectral gap $\rho = \Theta(M^{-2})$, there exist convex L-smooth functions $f_1, \dots, f_M : \mathbb{R}^d \to \mathbb{R}$ in dimension $d = \Theta(1 + S\sqrt{\rho})$ such that, for $f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x)$, the minimizer $x^\star$ satisfies $\|x^\star\| \leq R$, and every deterministic linear-span decentralized first-order method run for S gossip communication rounds has averaged output $\bar{x}_S$ satisfying*

$$f(\bar{x}_S) - f(x^\star) \geq cLR^2 \min\left\{1, \frac{1}{\rho S^2}\right\}.$$

The proof of Theorem 2 appears in Appendix F. The theorem is a communication lower bound: on the path, the hard instance forces useful coordinate information to alternate between two blocks separated by graph distance $\Theta(\rho^{-1/2})$, so each such alternation costs $\Theta(\rho^{-1/2})$ gossip rounds. In our distributed model over M machines, a budget of N local gradients implies $\Theta(N/M)$ gossip rounds, since each of the M nodes takes one sample and performs one gossip step per distributed round. Substituting $S = \Theta(N/M)$ in the asymptotic regime $S \gtrsim \rho^{-1/2}$ gives

$$\Omega\left(\frac{LR^2M^2}{\rho N^2}\right).$$

Thus, in the deterministic case, the network-dependent term in Theorem 1 is unimprovable up to logarithmic factors for deterministic linear-span decentralized methods, even on path networks. The stochastic term is optimal for an even simpler reason: even if communication were free, the problem would contain ordinary stochastic convex optimization with N stochastic-gradient samples, whose minimax error is $\Omega(\sigma R/\sqrt{N})$ [e.g., Agarwal et al., 2012]. Thus, the upper bound in Theorem 1 is optimal up to logarithmic factors (at least for linear-span decentralized methods whose only randomness comes from the stochastic gradients): the first term is forced by stochastic oracle noise, and the second term is forced by network-limited information propagation.

4.1 Proof sketch of Theorem 1

Algorithm 1 combines two Nesterov-type acceleration mechanisms: accelerated optimization of the iterates and accelerated gossip for reducing network disagreement. This combination is delicate because accelerated gradient methods are notoriously sensitive to persistent oracle errors: even small bias or inexactness in the gradient information can accumulate severely across iterations [Devolder et al., 2014]. In the decentralized setting such inexactness arises naturally, because limited communication means that the local gradients are evaluated at different local query points rather than at the exact network average. To use acceleration without losing its rate, the disagreement between the local query points must therefore be sufficiently small. This is incompatible with querying a newly formed accelerated point immediately: the point must first be mixed sufficiently across the network. The one-step-delayed accelerated method resolves this tension: it lets the algorithm sample gradients at query points that were already computed and communicated in previous rounds, while the current primal copy undergoes accelerated gossip to within sufficient mixing error for use in the next query.

One-step-delayed acceleration. The proof is built on a primal-dual recursion $z_{t+1} = z_t - (a_t/\beta)g_t$, $x_{t+1} = (1-\theta_t)x_t + \theta_t z_{t+1}$, initialized at $z_0 = x_0$, with weights $a_t > 0$, averaging parameters $\theta_t = a_t/A_t$ where $A_t = \sum_{i=0}^{t} a_i$, and regularization $\beta > 0$. By a standard telescoping argument (substituting $a_t g_t = \beta(z_t - z_{t+1})$ and using the three-point identity), combined with smoothness and convexity of f at a query point r_t, one obtains for any comparator u (later set to a minimizer $x^\star$) the estimate

$$A_{T-1}\big(f(x_T) - f(u)\big) \le \frac{\beta}{2}\|u - x_0\|^2 - \frac{\beta}{2}\sum_{t=0}^{T-1}\|z_{t+1} - z_t\|^2 + \frac{L}{2}\sum_{t=0}^{T-1} A_t \|x_{t+1} - r_t\|^2 . \tag{3}$$

The first term is a one-time distance penalty. The second, negative term, is a "stability budget" that can absorb positive errors; its presence is what makes acceleration work. The challenge is to choose the query point r_t so that the last sum is absorbed by this stability budget.

Standard acceleration (no delay). In the standard Nesterov scheme one queries at the momentum point

$$y_t := (1-\theta_t)x_t + \theta_t z_t = x_t + m_t(x_t - x_{t-1}),$$

where m_t is the corresponding momentum coefficient. Setting $r_t = y_t$ in Eq. (3), the distance from x_{t+1} to the query point simplifies to $x_{t+1} - y_t = \theta_t(z_{t+1} - z_t)$, so $\frac{LA_t}{2}\|x_{t+1} - y_t\|^2 = \frac{LA_t\theta_t^2}{2}\|z_{t+1} - z_t\|^2$. For the parameter choice $a_t = (t+1)/(128L)$, the coefficient $LA_t\theta_t^2/2$ is at most $\beta/2$, so each term in the positive sum is dominated by the corresponding term in the negative stability sum. This cancellation is what gives the accelerated $O(1/T^2)$ rate.

The delay problem and the compensated query. With a one-step delay, the gradient used at time t must have been queried *before* the current iterate x_t is known (because the gradient oracle returns its result one step late). Therefore, we cannot query at y_t, which depends on x_t. The naive fallback of querying at the stale momentum point y_{t-1} introduces a mismatch $y_t - y_{t-1}$ that contains a large inertial displacement proportional to $x_t - x_{t-1}$; this displacement is *not* proportional to $\theta_t(z_{t+1} - z_t)$ and hence cannot be absorbed by the stability budget.

Our key observation is that a *momentum-compensated* query point, defined for $t \ge 1$ by

$$q_t := x_{t-1} + m_{t-1}(1 + m_t)(x_{t-1} - x_{t-2}), \qquad q_0 := x_0,$$

resolves this problem. The extra factor $(1 + m_t)$ anticipates the inertial displacement that will appear when the next momentum point y_t is formed. Concretely, it is chosen so that the identity

$$y_t - q_t = (1 + m_t)(x_t - y_{t-1})$$

holds, which rewrites the delay mismatch entirely in terms of the iterate displacement $x_t - y_{t-1}$. Since $x_t - y_{t-1} = \theta_{t-1}(z_t - z_{t-1})$ by the primal-dual recursion, this displacement *is* controlled by the stability budget. Quantitatively, with the parameters of Theorem 3 (the one-step delayed accelerated rate proven in Appendix B),

$$\sum_{t=1}^{T-1} A_t \|y_t - q_t\|^2 \le 12 \sum_{t=0}^{T-2} A_t \|x_{t+1} - y_t\|^2 .$$

Since $x_{t+1} - q_t = (x_{t+1} - y_t) + (y_t - q_t)$, the total cost with query point $r_t = q_t$ in Eq. (3) is bounded by a constant multiple of $\sum_t A_t\theta_t^2\|z_{t+1} - z_t\|^2$, which is still absorbed by the negative stability term.

Restoring acceleration: the compensated query makes the delayed scheme pay essentially the same smoothness cost as the non-delayed one, so the accelerated $O(1/T^2)$ rate is preserved despite the one-step delay. The full stochastic and biased-oracle analysis in Theorem 3 starts from the same estimate Eq. (3) and adds error terms for stochastic noise $g_t - \nabla f(q_t)$ and residual disagreement bias.

Accelerated gossip. The second mechanism is accelerated gossip, which reduces inter-node disagreement exponentially fast while preserving the network average. The key observation is that, per coordinate, a single (lazy) gossip round $v \mapsto (1-\alpha_g)v + \alpha_g vP$ is exactly a gradient step with stepsize α_g on the quadratic disagreement objective $F(x) := \frac{1}{2}x(I-P)x^\mathsf{T}$, whose minimizers are the consensus vectors. On the disagreement subspace $\{x : xJ = 0\}$ (where $J := \mathbf{1}\mathbf{1}^\mathsf{T}/M$), F is ρ-strongly convex and 2-smooth, so applying Nesterov's accelerated gradient method yields exponential contraction at the optimal rate for this condition number. After B accelerated gossip rounds the Frobenius-norm disagreement of the $d \times M$ network state Y contracts as

$$\left\|Y^{(B)} - Y^{(B)}J\right\|_F \le \frac{2}{\sqrt{\rho}}\left(1 - \sqrt{\rho/2}\right)^{B/2}\left\|Y^{(0)} - Y^{(0)}J\right\|_F,$$

since right-multiplication by P acts on each row independently and preserves row averages. Solving for the number of rounds needed to reach disagreement ε gives the communication complexity $B = O(\rho^{-1/2}\log(\|Y^{(0)} - Y^{(0)}J\|_F/(\varepsilon\sqrt{\rho})))$. This is the $\rho^{-1/2}$ scaling of accelerated gossip, compared with ρ^{-1} for ordinary gossip; see Appendix C for the full derivation. In Theorem 1, B is set to make the residual disagreement small enough for the delayed accelerated estimate to hold.

Putting the pieces together. We now combine the one-step-delayed accelerated estimate and the accelerated gossip contraction to sketch how the two-term bound in Theorem 1 arises.

Dual use of each macro-round. Each of the T macro steps uses the same block of B micro-rounds for two concurrent purposes: (i) every worker evaluates B stochastic gradients at the already available delayed query point, forming a local minibatch of size B; and (ii) accelerated gossip runs for B communication rounds on the current primal copies. The total sample budget is therefore $N = MBT$.

Reduction to the single-node problem. Since accelerated gossip preserves network averages exactly, the averaged iterates $\bar{x}_t = (1/M)\sum_i x_{t,i}$ follow the one-step-delayed accelerated recursion from Eq. (3), driven by the averaged oracle $\widehat{G}_t = \frac{1}{M}\sum_{i=1}^M g_{t,i}$. Minibatching over the B gradient samples in each macro-round reduces the conditional variance of $\widehat{G}_t$ to at most $\sigma^2/(MB)$. The only remaining difference from a centralized gradient query at the network-average point $\bar{q}_t$ is a conditional bias caused by the nodes querying at slightly different local points $q_{t,i}$. Let $\mathcal{F}_t$ be the history at the start of macro step t, including Q_t. By L-smoothness of each f_i, this bias is bounded by

$$\left\|\mathbb{E}[\widehat{G}_t \mid \mathcal{F}_t] - \nabla f(\bar{q}_t)\right\| = \left\|\frac{1}{M}\sum_{i=1}^M\left(\nabla f_i(q_{t,i}) - \nabla f_i(\bar{q}_t)\right)\right\| \le \frac{L}{\sqrt{M}}\left\|Q_t - \bar{Q}_t\right\|_F,$$

where $Q_t - \bar{Q}_t$ measures the disagreement between the local query points and their average. The proof therefore controls a feedback loop: query disagreement creates oracle bias, while local gradient discrepancies create future disagreement. The full proof tracks these error terms jointly over the macro steps. The accelerated gossip length B is chosen large enough that this feedback is contractive, so the accumulated network-induced bias fits within the inexact-oracle tolerance of Theorem 3.

Deriving the convergence bound. Applying Theorem 3 to the averaged dynamics yields two terms. The *stochastic term* comes from the averaged-oracle variance $\sigma^2/(MB)$; after T macro steps it gives $\frac{\sigma R}{\sqrt{T}}\cdot\frac{1}{\sqrt{MB}} = \frac{\sigma R}{\sqrt{MBT}} = \frac{\sigma R}{\sqrt{N}}$, recovering the centralized statistical rate. The *deterministic (network) term* comes from the accelerated $O(1/T^2)$ optimization rate. Since $T = N/(MB)$ and $B = \widetilde{\Theta}(\rho^{-1/2})$, this term becomes $\frac{LR^2}{T^2} = \frac{LR^2M^2B^2}{N^2} = \widetilde{O}\left(\frac{LR^2M^2}{\rho N^2}\right)$.

Per-node guarantee. The above bound holds for the network-average iterate $\bar{x}_T$; the per-node bound follows from the final accelerated-gossip block and smoothness.

What this sketch omits. The full proof in Appendix D makes the following aspects precise, which are only outlined here: (i) the stochastic error analysis, including the exact dependence on

the minibatch variance and the interaction between stochastic noise and the delayed query; (ii) the bias-budget calculation showing that the gossip contraction makes the deterministic bias terms in Theorem 3 sum to the claimed network-dependent rate; (iii) the role of the heterogeneity parameter $\zeta_\star$, which enters only through the logarithmic factor Λ_N; and (iv) the precise numerical constants.

Discussion and limitations. Theorem 1 gives a rate that is optimal up to logarithmic factors within the stated linear-span class for decentralized stochastic smooth convex optimization in our synchronous, nonblocking gossip model on fixed undirected networks. In particular, its network-dependent term matches our lower bound in Theorem 2 on path networks up to these factors. These comparisons apply within the nonblocking model; other communication schedules may have different rates and tradeoffs. Parameter choices in Theorem 1 require knowledge of problem and network quantities, or valid bounds on them. Estimating these quantities and tuning the method can be difficult in practice, and attaining the same rate without such prior knowledge remains an open question. In addition, logarithmic factors and conservative constants may limit the gains under limited budgets.

5 Experimental evaluation

We evaluate the optimization performance of DDA-SGD and its scaling with the number of workers by comparing it with D-SGD [Koloskova et al., 2020] and DAT-SGD [Eisen et al., 2025] on a synthetic least-squares problem following the setup of Eisen et al. [2025]. Two variants of DDA-SGD are considered: one using the momentum schedule of Theorem 1, and a practical variant with constant momentum $m_t = \beta_g = 0.9$, $\theta_t = 0.1$, $\alpha_g = \frac{1}{2}$ and $\eta_t = 10\eta$, where η is the tuned stepsize. The experiments use ring and torus networks, whose gossip matrices are fixed and symmetric.

Experimental setup. We consider networks of $M \in \{4, 9, 25, 49, 100\}$ workers. Worker i has local objective $f_i(x) = \frac{1}{2}\|A_i x - b_i\|^2$, with Gaussian data of dimension $d = 50$. We vary the gradient noise variance σ^2 and heterogeneity ζ^2 over $\{1, 10\}$. Every method runs for 10^5 communication rounds, using one stochastic gradient per worker per round. For each method and configuration, we tune the stepsize to minimize final error, jointly with B for DDA-SGD. Results are averaged over three random seeds. Full details, results for all configurations, and block-length sensitivity appear in Appendix G.

				DDA-SGD	
Graph	M	D-SGD	DAT-SGD	constant	theorem
Ring	4	$4.8 \cdot 10^{-6}$	$1.3 \cdot 10^{-5}$	$\mathbf{3.1 \cdot 10^{-8}}$	$5.0 \cdot 10^{-7}$
	9	$3.0 \cdot 10^{-4}$	$6.4 \cdot 10^{-6}$	$\mathbf{4.7 \cdot 10^{-9}}$	$9.1 \cdot 10^{-8}$
	25	$6.1 \cdot 10^{-3}$	$6.8 \cdot 10^{-6}$	$\mathbf{1.0 \cdot 10^{-9}}$	$2.8 \cdot 10^{-8}$
	49	$1.9 \cdot 10^{-2}$	$3.0 \cdot 10^{-5}$	$\mathbf{6.3 \cdot 10^{-10}}$	$1.4 \cdot 10^{-8}$
	100	$4.3 \cdot 10^{-2}$	$6.7 \cdot 10^{-4}$	$9.2 \cdot 10^{-5}$	$\mathbf{6.0 \cdot 10^{-7}}$
Torus	4	$5.8 \cdot 10^{-6}$	$1.3 \cdot 10^{-5}$	$\mathbf{3.1 \cdot 10^{-8}}$	$5.0 \cdot 10^{-7}$
	9	$9.4 \cdot 10^{-6}$	$7.0 \cdot 10^{-6}$	$\mathbf{4.7 \cdot 10^{-9}}$	$9.1 \cdot 10^{-8}$
	25	$5.6 \cdot 10^{-5}$	$2.7 \cdot 10^{-6}$	$\mathbf{1.0 \cdot 10^{-9}}$	$2.8 \cdot 10^{-8}$
	49	$1.4 \cdot 10^{-4}$	$1.7 \cdot 10^{-6}$	$\mathbf{6.3 \cdot 10^{-10}}$	$1.4 \cdot 10^{-8}$
	100	$3.7 \cdot 10^{-4}$	$9.1 \cdot 10^{-7}$	$\mathbf{3.0 \cdot 10^{-10}}$	$8.1 \cdot 10^{-9}$

Table 2: Final error $\frac{1}{M}\sum_{i=1}^{M}\|x_i - x^\star\|^2$ on ring and torus networks for $\sigma^2 = \zeta^2 = 1$. Both DDA-SGD momentum schedules are shown.

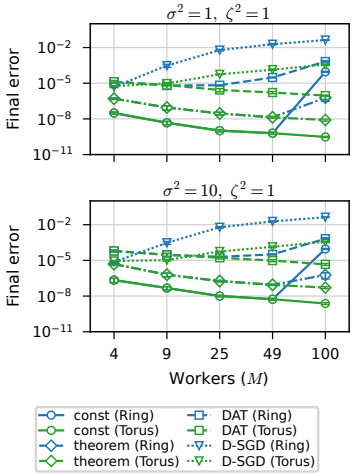

Figure 1: Final error versus the number of workers ($\zeta^2 = 1$).

Results. DDA-SGD achieves lower final error than both baselines across all tested configurations (Table 2). The constant and theorem momentum schedules improve on DAT-SGD by at least 6-fold and 12-fold, respectively. At a fixed per-worker budget, DDA-SGD benefits from additional workers through $M = 100$ on the torus and $M = 49$ on the ring (Fig. 1), whereas DAT-SGD's ring error is lowest at $M = 9$ or 25. On the 100-worker ring with $\sigma^2 = \zeta^2 = 1$, the theorem schedule substantially reduces the error relative to constant momenta: $6.0 \cdot 10^{-7}$ versus $9.2 \cdot 10^{-5}$.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 2549/19 and 3174/23), from the Council for Higher Education in Israel under a Moonshot Project, a grant from the Tel Aviv University Center for AI and Data Science (TAD), from the Len Blavatnik and the Blavatnik Family foundation, and a fellowship from the Israeli Council for Higher Education.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pages 344–353. PMLR, 2019.

Amit Attia, Ofir Gaash, and Tomer Koren. Faster stochastic optimization with arbitrary delays via adaptive asynchronous mini-batching. In *Forty-second International Conference on Machine Learning*, 2025.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, et al. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Ofri Eisen, Ron Dorfman, and Kfir Yehuda Levy. Enhancing parallelism in decentralized stochastic convex optimization. In *International Conference on Machine Learning*, pages 15105–15129. PMLR, 2025.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippedgossip. *arXiv preprint arXiv:2202.01545*, 2022.

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 482–491. IEEE, 2003.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pages 5381–5393. PMLR, 2020.

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34: 11422–11435, 2021.

Dmitry Kovalev, Egor Shulgin, Peter Richtárik, Alexander V Rogozin, and Alexander Gasnikov. Adom: accelerated decentralized optimization method for time-varying networks. In *International Conference on Machine Learning*, pages 5784–5793. PMLR, 2021.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication-efficient local decentralized SGD methods. *arXiv preprint arXiv:1910.09126*, 2019.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pages 7111–7123. PMLR, 2021.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on automatic control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2 edition, 2018.

Alexander Rogozin, César A Uribe, Alexander V Gasnikov, Nikolay Malkovsky, and Angelia Nedic. Optimal distributed convex optimization on slowly time-varying graphs. *IEEE Trans. Control. Netw. Syst.*, 7(2):829–841, 2020.

Alexander Rogozin, Vladislav Lukoshkin, Alexander Gasnikov, Dmitry Kovalev, and Egor Shulgin. Towards accelerated rates for distributed optimization over time-varying networks. In *International Conference on Optimization and Applications*, pages 258–272. Springer, 2021.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011. ISSN 1935-8237. doi: 10.1561/2200000018.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018b.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9): 803–812, 1986.

Alexander Tyurin and Peter Richtárik. Optimal time complexities of parallel stochastic optimization methods under a fixed computation model. *Advances in Neural Information Processing Systems*, 36:16515–16577, 2023.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks. In *International Conference on Artificial Intelligence and Statistics*, pages 2381–2391. PMLR, 2020.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of machine learning research*, 24(306):1–50, 2023.

A Additional related work

Gossip algorithms and consensus. Gossip protocols have a long history in distributed computing, originating with the early work of Tsitsiklis et al. [1986] on distributed asynchronous gradient algorithms, which established that consensus-based gradient methods retain the convergence properties of their centralized counterparts under bounded communication delays. Randomized gossip averaging was formalized by Kempe et al. [2003], who showed that simple push-sum protocols compute aggregate statistics in $O(\log n)$ rounds on general networks, and by Boyd et al. [2006], who characterized the averaging time of pairwise randomized gossip in terms of the second-largest eigenvalue of the gossip matrix and showed that optimal weight design reduces to a semidefinite program. Xiao and Boyd [2004] studied the fastest linear iterations for distributed averaging and demonstrated that optimally designed gossip matrices substantially outperform Laplacian-based heuristics. This spectral-gap perspective, linking the mixing rate of the gossip matrix to the convergence of distributed algorithms, is central to the analysis of all subsequent gossip-based optimization methods, including ours.

B One-step-delayed accelerated algorithm

In this section we isolate the one-step-delayed acceleration component of our gossip method. Through minibatching and accelerated gossip, the network average reduces to a single-node stochastic optimization problem with a delayed oracle, while residual disagreement enters as a controlled oracle bias. We present an accelerated algorithm for this reduced problem and prove that momentum-compensated queries retain the usual stochastic accelerated rate.

B.1 Problem setup

We consider the convex stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is convex and differentiable. We assume that f has an L-Lipschitz gradient:

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|, \qquad x,y\in\mathbb{R}^d.$$

Equivalently,

$$f(u)\le f(v)+\langle\nabla f(v),u-v\rangle+\frac{L}{2}\|u-v\|^2.$$

Let $x^\star\in\arg\min_{x\in\mathbb{R}^d} f(x)$, and assume that $R>0$ satisfies

$$R\ge\left\|x_0-x^\star\right\|.$$

If the minimizer is not unique, one may take $x^\star$ to be any minimizer and R to be any known upper bound on its distance from x_0; the sharpest such choice is the distance from x_0 to the solution set.

Delayed stochastic oracle. At round t, the algorithm receives a stochastic gradient queried at the previous round. If the point queried at the previous round is q_t, then the returned random vector is

$$G_t=G(q_t,\xi_t)=\nabla f(q_t)+\zeta_t.$$

The stochastic error satisfies the RMS conditional-bias and centered-variance assumptions

$$b_t:=\mathbb{E}[\zeta_t\mid\mathcal{F}_t],\qquad \mathbb{E}\|b_t\|^2\le\epsilon_t^2,\qquad \mathbb{E}[\|\zeta_t-b_t\|^2\mid\mathcal{F}_t]\le\sigma^2, \tag{4}$$

where $\epsilon_0,\dots,\epsilon_{T-1}$ are deterministic nonnegative RMS bias bounds for the horizon under consideration, and $\mathcal{F}_t$ contains all randomness and all iterates available before G_t is revealed. In particular, q_t is $\mathcal{F}_t$-measurable. The unbiased model is recovered by taking $\epsilon_t=0$ for all t. In the unconstrained setting we also assume that the oracle calls made by the algorithm are square-integrable:

$$\mathbb{E}\|G(q_t,\xi_t)\|^2<\infty,\qquad 0\le t<T. \tag{5}$$

B.2 Algorithm

The algorithm is written for prescribed weights a_t, averaging parameters θ_t, momentum parameters m_t, and prox coefficient β. The concrete parameter values used in the convergence theorem are stated in Theorem 3.

At round t, the next query point q_{t+1} is already computable from the pre-gradient state (x_t, x_{t-1}) and the known momentum coefficients; it does not depend on the delayed gradient G_t. This is what makes the method compatible with a strict one-step delay. The method then receives the stochastic gradient of the point submitted one round earlier, takes the dual gradient step $z_{t+1} = z_t - (a_t/\beta)G_t$, and forms the accelerated primal iterate x_{t+1} by averaging x_t with z_{t+1}. This recursion is equivalent to the cumulative prox/FTRL (Follow-the-Regularized-Leader) update $z_{t+1} = x_0 - \beta^{-1}\sum_{i=0}^{t} a_i G_i$, minimizing cumulative linear losses plus a quadratic regularizer.

Algorithm 2 One-step delayed momentum-compensated accelerated stochastic approximation

1: Input: $x_0 \in \mathbb{R}^d$, horizon T, parameters $(a_t, \theta_t)_{t=0}^{T-1}$, $(m_t)_{t=0}^{T}$, and $\beta > 0$.
2: Set $z_0 = x_0$, $x_{-1} = x_0$, and $q_0 = x_0$.
3: Submit the initial query at q_0.
4: **for** $t = 0, 1, \ldots, T-1$ **do**
5: Compute the next compensated query point

$$q_{t+1} = x_t + m_t(1 + m_{t+1})(x_t - x_{t-1}),$$

with $m_0 = 0$.
6: Submit the next-round query at q_{t+1}; receive the delayed stochastic gradient $G_t = G(q_t, \xi_t)$ from the previously submitted query at q_t.
7: Take the dual gradient step

$$z_{t+1} = z_t - \frac{a_t}{\beta} G_t.$$

8: Compute the accelerated primal point

$$x_{t+1} = (1 - \theta_t)x_t + \theta_t z_{t+1}.$$

9: **end for**
10: Output x_T.

Compensated query points. For a Nesterov-type momentum sequence, the extrapolated point is

$$y_t = (1 - \theta_t)x_t + \theta_t z_t = x_t + m_t(x_t - x_{t-1}).$$

If the gradient at round t must have been queried at round $t-1$, querying the naive stale point y_{t-1} leads to the unstable delayed-Nesterov recurrence. Instead, Algorithm 2 queries the one-step momentum-compensated point

$$q_t := x_{t-1} + m_{t-1}(1 + m_t)(x_{t-1} - x_{t-2}), \qquad t \geq 1, \tag{6}$$

with the convention $q_0 = x_0$.

The extra term

$$m_{t-1}m_t(x_{t-1} - x_{t-2})$$

is the one-step momentum compensation. It anticipates the additional inertial displacement that will appear between y_{t-1} and y_t.

B.3 Main result

Theorem 3 (Optimal one-step delayed stochastic accelerated rate)**.** *Assume Eqs.* (4) *and* (5) *hold. Fix a horizon $T \geq 1$, and set*

$$a_t = \frac{t+1}{128L}, \qquad A_t = \sum_{i=0}^{t} a_i = \frac{(t+1)(t+2)}{256L}, \qquad \theta_t = \frac{a_t}{A_t} = \frac{2}{t+2}, \qquad 0 \leq t \leq T.$$

Set

$$m_0 = 0, \qquad m_t := \frac{\theta_t(1-\theta_{t-1})}{\theta_{t-1}} = \frac{t-1}{t+2}, \qquad 1 \le t \le T.$$

Finally, define

$$H_{T-1} := \sqrt{\sum_{i=0}^{T-1} a_i^2}, \qquad \beta := 1 + \frac{\sigma}{R} H_{T-1}. \tag{7}$$

Run Algorithm 2 with these parameters. Then

$$\mathbb{E}[f(x_T) - f(x^\star)] \le \frac{6\sigma R}{\sqrt{T}} + \frac{128LR^2}{T^2} + \mathcal{E}_T, \tag{8}$$

where

$$\mathcal{E}_T := 2^{11}\left(R + \sqrt{\frac{\sigma R}{L}}\, T^{3/4}\right) T\Gamma_T + \frac{64}{L} T^4 \Gamma_T^2,$$

and Γ_T is the weighted average RMS bias,

$$\Gamma_T := \frac{1}{A_{T-1}} \sum_{t=0}^{T-1} a_t \epsilon_t. \tag{9}$$

If $\epsilon_t = 0$ for all t, then $\mathcal{E}_T = 0$ and the last term in Eq. (8) *vanishes.*

Relation to inexact-oracle acceleration. The bias dependence in Theorem 3 is consistent with the inexact-oracle analysis of Devolder, Glineur, and Nesterov [2014]. Their accelerated bounds for a scalar (δ, L)-inexact oracle contain an accumulated error term of order $T\delta$. In the deterministic case $\sigma = 0$, the leading bias term above is $RT\Gamma_T$, which has the same form after interpreting the scalar oracle error as $\delta \sim R\Gamma_T$. The additional linear term involving $\sqrt{\sigma R/L}\, T^{1+3/4}\Gamma_T$ comes from stochastic spread of the iterates. The final quadratic term $(1/L)T^4\Gamma_T^2$ is not a feature of the scalar inexact-oracle model itself; it appears here because we control vector-valued bias on an unconstrained domain. The proof must first bound the RMS size of $z_t - x^\star$ in terms of the accumulated bias, and this self-consistency step produces the quadratic drift contribution. Thus the Devolder et al. [2014]-type $O(TR\Gamma_T)$ accumulation is the leading effect, while the quadratic term is the price of converting vector bias into scalar function error without an a priori radius bound.

B.4 Proofs

B.4.1 Elementary identities

The following identity is the algebraic reason for the compensation. Without the compensated query point in (cf. Eq. (6)), the stale point would be y_{t-1}, and the difference $y_t - y_{t-1}$ contains an uncontrolled inertial component. The scalar quadratic test then produces an unstable limiting recurrence. The compensated point removes precisely the hard inertial component. Lemma 1 below shows that the remaining mismatch is proportional only to the previous actual update $x_t - y_{t-1}$, which is controlled by the negative stability terms in the accelerated estimate.

Lemma 1 (One-step compensation identity)**.** *For the parameter choice in Theorem 3 and every $t \ge 1$, Algorithm 2 satisfies*

$$y_t - q_t = (1 + m_t)(x_t - y_{t-1}). \tag{10}$$

Consequently,

$$\|y_t - q_t\|^2 \le 4\,\|x_t - y_{t-1}\|^2. \tag{11}$$

Proof. Since

$$y_s = x_s + m_s(x_s - x_{s-1}),$$

we have

$$x_t - y_{t-1} = x_t - x_{t-1} - m_{t-1}(x_{t-1} - x_{t-2}).$$

Using the definition of q_t,

$$\begin{aligned} y_t - q_t &= x_t + m_t(x_t - x_{t-1}) - x_{t-1} - m_{t-1}(1 + m_t)(x_{t-1} - x_{t-2}) \\ &= (1 + m_t)(x_t - x_{t-1}) - (1 + m_t)m_{t-1}(x_{t-1} - x_{t-2}) \\ &= (1 + m_t)(x_t - y_{t-1}). \end{aligned}$$

Since $0 \le m_t < 1$, $1 + m_t \le 2$, giving Eq. (11). □

Lemma 2 (Weighted delay-energy bound)**.** *For the parameter choice in Theorem 3 and every $T \geq 2$,*

$$\sum_{t=1}^{T-1} A_t \left\|y_t - q_t\right\|^2 \leq 12 \sum_{t=0}^{T-2} A_t \left\|x_{t+1} - y_t\right\|^2. \tag{12}$$

Proof. By Lemma 1,

$$\left\|y_t - q_t\right\|^2 \leq 4 \left\|x_t - y_{t-1}\right\|^2.$$

For $t \geq 1$,

$$\frac{A_t}{A_{t-1}} = \frac{(t+1)(t+2)}{t(t+1)} = \frac{t+2}{t} \leq 3.$$

Therefore

$$A_t \left\|y_t - q_t\right\|^2 \leq 12 A_{t-1} \left\|x_t - y_{t-1}\right\|^2.$$

Summing over $t = 1, \ldots, T-1$ and re-indexing gives Eq. (12). □

B.4.2 Deterministic accelerated estimate with delayed gradients

We first state the deterministic estimate-sequence inequality that will be used inside the stochastic proof. Its role is to show that the compensated stale query behaves like an ordinary accelerated gradient step with a larger smoothness constant.

Lemma 3 (Compensated delayed accelerated estimate)**.** *Let $T \geq 1$. Let g_t be any sequence of vectors, and define*

$$\delta_t := g_t - \nabla f(q_t).$$

Suppose the iterates are generated by Algorithm 2 with the parameter choice in Theorem 3. Then, for every $u \in \mathbb{R}^d$,

$$A_{T-1}\big(f(x_T) - f(u)\big) + 3L \sum_{t=0}^{T-1} A_t \left\|x_{t+1} - y_t\right\|^2 \leq \frac{\beta}{2}\left\|u - x_0\right\|^2 + \sum_{t=0}^{T-1} a_t \left\langle \delta_t, u - z_t \right\rangle + \sum_{t=0}^{T-1} \frac{a_t^2}{\beta}\left\|\delta_t\right\|^2. \tag{13}$$

Proof. We give the proof in three steps.

Step 1: a smoothness inequality at the compensated point. Set $A_{-1} := 0$. For each t, let

$$y_t = (1-\theta_t)x_t + \theta_t z_t, \qquad x_{t+1} = (1-\theta_t)x_t + \theta_t z_{t+1}.$$

Smoothness at the query point q_t gives

$$f(x_{t+1}) \leq f(q_t) + \langle \nabla f(q_t), x_{t+1} - q_t \rangle + \frac{L}{2}\left\|x_{t+1} - q_t\right\|^2.$$

By convexity, applied at q_t and then averaged between x_t and u,

$$(1-\theta_t) f(x_t) + \theta_t f(u) \geq f(q_t) + \langle \nabla f(q_t), (1-\theta_t)x_t + \theta_t u - q_t \rangle.$$

Since $x_{t+1} = (1-\theta_t)x_t + \theta_t z_{t+1}$, the last two displays imply

$$A_t f(x_{t+1}) - A_{t-1} f(x_t) \leq a_t f(u) + a_t \langle \nabla f(q_t), z_{t+1} - u \rangle + \frac{L A_t}{2}\left\|x_{t+1} - q_t\right\|^2. \tag{14}$$

Step 2: the fixed-prox FTRL inequality. The following argument establishes the FTRL inequality for a fixed quadratic regularizer centered at x_0; see Shalev-Shwartz [2011], Hazan [2016] for background on FTRL. Let $s_t := a_t g_t$ and $S_t := \sum_{i=0}^{t} s_i$. For $t \geq 0$, define

$$h_t(z) := \langle S_t, z \rangle + \frac{\beta}{2}\left\|z - x_0\right\|^2, \qquad z \in \mathbb{R}^d,$$

and set $h_{-1}(z) := \frac{\beta}{2}\left\|z - x_0\right\|^2$. The recursion $z_{t+1} = z_t - s_t/\beta$ and $z_0 = x_0$ imply $z_{t+1} = x_0 - S_t/\beta$. Thus z_{t+1} minimizes h_t on $\mathbb{R}^d$, and $z_0 = x_0$ minimizes h_{-1} on $\mathbb{R}^d$.

Because each h_t is β-strongly convex on $\mathbb{R}^d$,

$$h_{t-1}(z_t) + \frac{\beta}{2}\left\|z_{t+1} - z_t\right\|^2 \leq h_{t-1}(z_{t+1}).$$

Using $h_t(z_{t+1}) - h_{t-1}(z_{t+1}) = \langle s_t, z_{t+1} \rangle$, we obtain

$$\langle s_t, z_{t+1} \rangle \le h_t(z_{t+1}) - h_{t-1}(z_t) - \frac{\beta}{2} \|z_{t+1} - z_t\|^2 .$$

Summing over $t = 0, \dots, T-1$ telescopes:

$$\sum_{t=0}^{T-1} \langle s_t, z_{t+1} \rangle \le h_{T-1}(z_T) - h_{-1}(z_0) - \frac{\beta}{2} \sum_{t=0}^{T-1} \|z_{t+1} - z_t\|^2 .$$

Since z_T minimizes h_{T-1} and $h_{-1}(z_0) = 0$, for every $u \in \mathbb{R}^d$,

$$\sum_{t=0}^{T-1} a_t \langle g_t, z_{t+1} - u \rangle \le \frac{\beta}{2} \|u - x_0\|^2 - \frac{\beta}{2} \sum_{t=0}^{T-1} \|z_{t+1} - z_t\|^2 . \tag{15}$$

Step 3: summation and absorption of delay energy. Substitute $\nabla f(q_t) = g_t - \delta_t$ into Eq. (14), sum over $t = 0, \dots, T-1$, and use Eq. (15). This gives

$$\begin{aligned} A_{T-1}\big(f(x_T) - f(u)\big) \le{}& \frac{\beta}{2} \|u - x_0\|^2 + \sum_{t=0}^{T-1} a_t \langle \delta_t, u - z_{t+1} \rangle \\ & - \frac{\beta}{2} \sum_{t=0}^{T-1} \|z_{t+1} - z_t\|^2 + \frac{L}{2} \sum_{t=0}^{T-1} A_t \|x_{t+1} - q_t\|^2 . \end{aligned}$$

Decompose

$$\langle \delta_t, u - z_{t+1} \rangle = \langle \delta_t, u - z_t \rangle + \langle \delta_t, z_t - z_{t+1} \rangle .$$

Young's inequality with parameter $\beta/2$ gives

$$a_t \langle \delta_t, z_t - z_{t+1} \rangle \le \frac{a_t^2}{\beta} \|\delta_t\|^2 + \frac{\beta}{4} \|z_{t+1} - z_t\|^2 .$$

Thus only the following residual deterministic term remains to be controlled:

$$\frac{L}{2} \sum_{t=0}^{T-1} A_t \|x_{t+1} - q_t\|^2 - \frac{\beta}{4} \sum_{t=0}^{T-1} \|z_{t+1} - z_t\|^2 .$$

Since $x_{t+1} - y_t = \theta_t (z_{t+1} - z_t)$ and $\beta \ge 1$,

$$\frac{\beta}{4} \|z_{t+1} - z_t\|^2 = \frac{\beta}{4\theta_t^2} \|x_{t+1} - y_t\|^2 \ge 16 L A_t \|x_{t+1} - y_t\|^2 ,$$

where the last inequality follows from

$$\frac{1}{\theta_t^2} = \frac{(t+2)^2}{4} \ge 64 L A_t = \frac{(t+1)(t+2)}{4}.$$

Moreover, $q_0 = y_0 = x_0$ and

$$\|x_{t+1} - q_t\|^2 \le 2 \|x_{t+1} - y_t\|^2 + 2 \|y_t - q_t\|^2 .$$

Using Lemma 2 when $T \ge 2$ and the trivial empty-sum bound when $T = 1$,

$$\frac{L}{2} \sum_{t=0}^{T-1} A_t \|x_{t+1} - q_t\|^2 \le 13 L \sum_{t=0}^{T-1} A_t \|x_{t+1} - y_t\|^2 .$$

Because $16L - 13L = 3L$, the residual deterministic term is bounded above by

$$-3L \sum_{t=0}^{T-1} A_t \|x_{t+1} - y_t\|^2 .$$

Moving this term to the left proves Eq. (13). □

B.4.3 A crude RMS-distance bound

Lemma 4 (Uniform RMS distance of query and dual points)**.** *Under the assumptions and parameter choice of Theorem 3, define*

$$B_T^{\epsilon} := 1024(T+1)\left(R + \sqrt{\frac{\sigma R}{L}}\, T^{3/4} + (T+1)A_{T-1}\Gamma_T\right).$$

Then, for every $0 \le t \le T$*,*

$$\left(\mathbb{E}\left\|q_t - x^\star\right\|^2\right)^{1/2} \le B_T^{\epsilon}, \qquad \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2} \le B_T^{\epsilon}.$$

Proof. Let

$$d_t := x_{t+1} - y_t.$$

Set

$$S := \sum_{t=0}^{T-1} A_t \mathbb{E}\left\|d_t\right\|^2, \qquad D_t := \sum_{i=0}^{t-1} \left\|d_i\right\|.$$

For $0 \le s \le T$, set $v_s := x_s - x_{s-1}$, with $v_0 = 0$. For $0 \le s \le T-1$, the identity

$$x_{s+1} = y_s + d_s = x_s + m_s(x_s - x_{s-1}) + d_s$$

implies

$$v_{s+1} = m_s v_s + d_s.$$

Since $0 \le m_s < 1$, induction gives

$$\left\|v_s\right\| \le \sum_{i=0}^{s-1} \left\|d_i\right\| \le D_s, \qquad s \ge 1.$$

Consequently, for $s \ge 0$,

$$\left\|x_s - x^\star\right\| \le R + sD_s.$$

For $t \ge 1$, using

$$q_t = x_{t-1} + m_{t-1}(1+m_t)v_{t-1}$$

and $m_{t-1}(1+m_t) \le 2$ gives

$$\left\|q_t - x^\star\right\| \le \left\|x_{t-1} - x^\star\right\| + 2\left\|v_{t-1}\right\| \le R + (t+1)D_t.$$

Also, for $t \ge 1$,

$$x_t = (1-\theta_{t-1})x_{t-1} + \theta_{t-1} z_t,$$

so

$$z_t = x_{t-1} + \frac{v_t}{\theta_{t-1}}.$$

Since $\theta_{t-1} = 2/(t+1)$,

$$\left\|z_t - x^\star\right\| \le \left\|x_{t-1} - x^\star\right\| + \frac{t+1}{2}\left\|v_t\right\| \le R + \left(t - 1 + \frac{t+1}{2}\right) D_t \le R + 2(t+1)D_t.$$

The same two bounds hold at $t = 0$ because $q_0 = z_0 = x_0$.

By Cauchy-Schwarz inequality and the explicit formula for A_i,

$$D_t^2 \le \left(\sum_{i=0}^{t-1} \frac{1}{A_i}\right)\left(\sum_{i=0}^{t-1} A_i \left\|d_i\right\|^2\right),$$

and

$$\sum_{i=0}^{t-1} \frac{1}{A_i} = 256L \sum_{i=0}^{t-1} \frac{1}{(i+1)(i+2)} \le 256L.$$

Therefore

$$\left(\mathbb{E}D_T^2\right)^{1/2} \le 16\sqrt{LS}.$$

We next bound S. Apply Lemma 3 with $u = x^\star$ and $\delta_t = \zeta_t$, take expectations, and drop the nonnegative term $A_{T-1}\mathbb{E}[f(x_T) - f(x^\star)]$, giving

$$3LS \le \frac{\beta R^2}{2} + \sum_{t=0}^{T-1} a_t \mathbb{E}\left\langle \zeta_t, x^\star - z_t \right\rangle + \frac{1}{\beta}\sum_{t=0}^{T-1} a_t^2 \mathbb{E}\left\|\zeta_t\right\|^2.$$

The square-integrability assumption and the recursion for z_t imply that each z_t is square-integrable. Since z_t is also $\mathcal{F}_t$-measurable and $b_t = \mathbb{E}[\zeta_t \mid \mathcal{F}_t]$, this gives

$$3LS \le \frac{\beta R^2}{2} + \frac{\sigma^2}{\beta} H_{T-1}^2 + \frac{1}{\beta}\sum_{t=0}^{T-1} a_t^2 \epsilon_t^2 + \sum_{t=0}^{T-1} a_t \epsilon_t \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2}.$$

Here we used the orthogonality $\mathbb{E}\left\langle \zeta_t - b_t, b_t\right\rangle = 0$, which follows from $b_t = \mathbb{E}[\zeta_t \mid \mathcal{F}_t]$, to get

$$\mathbb{E}\left\|\zeta_t\right\|^2 = \mathbb{E}\left\|\zeta_t - b_t\right\|^2 + \mathbb{E}\left\|b_t\right\|^2 \le \sigma^2 + \epsilon_t^2$$

and Cauchy-Schwarz inequality in expectation:

$$\mathbb{E}\left\langle b_t, x^\star - z_t\right\rangle \le \left(\mathbb{E}\left\|b_t\right\|^2\right)^{1/2}\left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2}.$$

Using the pathwise bound on z_t and the preceding estimate on $\left(\mathbb{E}D_T^2\right)^{1/2}$,

$$\sum_{t=0}^{T-1} a_t \epsilon_t \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2} \le A_{T-1}\Gamma_T\left(R + 32(T+1)\sqrt{LS}\right).$$

Also,

$$\frac{1}{\beta}\sum_{t=0}^{T-1} a_t^2\epsilon_t^2 \le \left(\sum_{t=0}^{T-1} a_t\epsilon_t\right)^2 = A_{T-1}^2\Gamma_T^2.$$

Moreover,

$$\frac{\beta R^2}{2} + \frac{\sigma^2}{\beta}H_{T-1}^2 \le \frac{R^2 + 3\sigma R H_{T-1}}{2},$$

because $\beta = 1 + \sigma H_{T-1}/R$ and $\sigma H_{T-1}/\beta \le R$. With $Y := \sqrt{LS}$, we have

$$3Y^2 \le \frac{R^2 + 3\sigma R H_{T-1}}{2} + A_{T-1}^2\Gamma_T^2 + RA_{T-1}\Gamma_T + 32(T+1)A_{T-1}\Gamma_T Y.$$

The quadratic inequality yields

$$Y \le \frac{32}{3}(T+1)A_{T-1}\Gamma_T + \sqrt{\frac{R^2 + 3\sigma R H_{T-1} + 2RA_{T-1}\Gamma_T + 2A_{T-1}^2\Gamma_T^2}{6}}.$$

We simplify this bound using

$$\sqrt{x_1 + \cdots + x_4} \le \sqrt{x_1} + \cdots + \sqrt{x_4}, \qquad \frac{32}{3} \le 16, \qquad H_{T-1} \le \frac{T^{3/2}}{64\sqrt{3}L},$$

and

$$\sqrt{RA_{T-1}\Gamma_T} \le R + (T+1)A_{T-1}\Gamma_T.$$

We also use $A_{T-1}\Gamma_T \le (T+1)A_{T-1}\Gamma_T$. This gives

$$Y \le 20\left(R + \sqrt{\frac{\sigma R}{L}}\,T^{3/4} + (T+1)A_{T-1}\Gamma_T\right).$$

Combining this with the bounds

$$\left(\mathbb{E}\left\|q_t - x^\star\right\|^2\right)^{1/2} \le R + 16(T+1)Y, \qquad \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2} \le R + 32(T+1)Y$$

and using $T \ge 1$ proves the lemma by the definition of B_T^ϵ. □

B.4.4 Proof of Theorem 3

Apply Lemma 3 with $u = x^\star$ and $\delta_t = \zeta_t$, and drop the nonnegative movement term. We obtain

$$A_{T-1}\big(f(x_T) - f(x^\star)\big) \le \frac{\beta}{2}\left\|x^\star - x_0\right\|^2 + \sum_{t=0}^{T-1} a_t \left\langle \zeta_t, x^\star - z_t \right\rangle + \sum_{t=0}^{T-1} \frac{a_t^2}{\beta}\left\|\zeta_t\right\|^2.$$

The square-integrability assumption implies by induction that each $z_t = x_0 - \beta^{-1}\sum_{i=0}^{t-1} a_i G_i$ is square-integrable. Hence the inner products below are integrable. Because z_t is $\mathcal{F}_t$-measurable and $b_t = \mathbb{E}[\zeta_t \mid \mathcal{F}_t]$,

$$\mathbb{E}\left[a_t \left\langle \zeta_t, x^\star - z_t \right\rangle\right] = \mathbb{E}\left[a_t \left\langle b_t, x^\star - z_t \right\rangle\right] \le a_t \epsilon_t \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2}.$$

By Lemma 4,

$$\sum_{t=0}^{T-1} a_t \epsilon_t \left(\mathbb{E}\left\|z_t - x^\star\right\|^2\right)^{1/2} \le A_{T-1}\Gamma_T B_T^\epsilon.$$

Also, since $b_t = \mathbb{E}[\zeta_t \mid \mathcal{F}_t]$, the same orthogonality gives

$$\mathbb{E}\left\|\zeta_t\right\|^2 = \mathbb{E}\left\|\zeta_t - b_t\right\|^2 + \mathbb{E}\left\|b_t\right\|^2 \le \sigma^2 + \epsilon_t^2.$$

Therefore

$$\mathbb{E}[f(x_T) - f(x^\star)] \le \frac{\beta R^2}{2A_{T-1}} + \frac{\sigma^2}{A_{T-1}\beta}\sum_{t=0}^{T-1} a_t^2 + \frac{1}{A_{T-1}\beta}\sum_{t=0}^{T-1} a_t^2\epsilon_t^2 + \Gamma_T B_T^\epsilon. \tag{16}$$

By definition,

$$\beta = 1 + \frac{\sigma}{R}H_{T-1}.$$

Thus

$$\frac{\sigma^2}{\beta}\sum_{t=0}^{T-1} a_t^2 = \frac{\sigma^2}{\beta}H_{T-1}^2 \le \sigma R H_{T-1},$$

because $\sigma H_{T-1}/\beta \le R$. Also,

$$\frac{1}{\beta}\sum_{t=0}^{T-1} a_t^2\epsilon_t^2 \le \left(\sum_{t=0}^{T-1} a_t\epsilon_t\right)^2 = A_{T-1}^2\Gamma_T^2.$$

Substituting into Eq. (16) gives

$$\mathbb{E}[f(x_T) - f(x^\star)] \le \frac{R^2}{2A_{T-1}} + \frac{\sigma R H_{T-1}}{2A_{T-1}} + \frac{\sigma R H_{T-1}}{A_{T-1}} + A_{T-1}\Gamma_T^2 + \Gamma_T B_T^\epsilon.$$

Hence

$$\mathbb{E}[f(x_T) - f(x^\star)] \le \frac{R^2}{2A_{T-1}} + \frac{3\sigma R H_{T-1}}{2A_{T-1}} + A_{T-1}\Gamma_T^2 + \Gamma_T B_T^\epsilon. \tag{17}$$

Now compute

$$A_{T-1} = \frac{T(T+1)}{256L} \ge \frac{T^2}{256L},$$

and

$$H_{T-1}^2 = \sum_{t=0}^{T-1}\frac{(t+1)^2}{16384L^2} = \frac{T(T+1)(2T+1)}{98304L^2} \le \frac{T^3}{12288L^2}.$$

Therefore

$$\frac{R^2}{2A_{T-1}} = \frac{128LR^2}{T(T+1)},$$

and

$$\frac{3\sigma R H_{T-1}}{2A_{T-1}} \le \frac{3\sigma R}{2}\cdot\frac{T^{3/2}/(64\sqrt{3}L)}{T(T+1)/(256L)} \le \frac{6\sigma R}{\sqrt{3T}}.$$

Finally, by the definitions of B_T^ϵ and A_{T-1},

$$\begin{aligned} A_{T-1}\Gamma_T^2 + \Gamma_T B_T^\epsilon &= 2^{10}(T+1)\Gamma_T \left(R + \sqrt{\frac{\sigma R}{L}}\, T^{3/4} + (T+1)A_{T-1}\Gamma_T \right) + A_{T-1}\Gamma_T^2 \\ &= 2^{10}(T+1)\Gamma_T \left(R + \sqrt{\frac{\sigma R}{L}}\, T^{3/4} \right) + \left(\frac{4T(T+1)^3}{L} + \frac{T(T+1)}{256L} \right) \Gamma_T^2 \\ &\le 2^{11} T \Gamma_T \left(R + \sqrt{\frac{\sigma R}{L}}\, T^{3/4} \right) + \frac{64}{L} T^4 \Gamma_T^2. \end{aligned}$$

Combining these estimates proves Eq. (8). □

C Accelerated gossip via Nesterov acceleration

This appendix records the accelerated communication primitive used in the distributed algorithm. The standard way to accelerate gossip is to apply a Chebyshev polynomial in the gossip matrix, which improves the communication round complexity from ρ^{-1} to $\rho^{-1/2}$. Here we give, for completeness, an alternative derivation from an optimization viewpoint: the accelerated version below applies Nesterov's method to a quadratic disagreement objective, giving the same $\rho^{-1/2}$ dependence while preserving the network average exactly. We present the argument for scalar local variables; matrix gossip in the main algorithm follows by applying the same recursion row by row.

Setup. Let $P \in \mathbb{R}^{n\times n}$ be a symmetric doubly-stochastic gossip matrix on a connected graph. In this section each node holds a scalar, so the global state is a row vector $x \in \mathbb{R}^{1\times n}$, identified with $\mathbb{R}^n$. Gossip acts by right multiplication, matching the convention used in the main algorithm. When the main algorithm gossips a matrix of vector-valued local copies, the vector result below is applied independently to each row, giving the corresponding Frobenius-norm contraction. Let

$$J := \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}.$$

Thus, for an initial row vector $\xi \in \mathbb{R}^{1\times n}$, the vector ξJ has all coordinates equal to the average of the coordinates of ξ. Write the eigenvalues of P as $1 = \lambda_1(P) > \lambda_2(P) \ge \cdots \ge \lambda_n(P)$, and assume that the spectral gap satisfies $1 - \lambda_2(P) \ge \rho$. Consider the quadratic disagreement objective

$$F(x) := \frac{1}{2} x(I-P)x^\mathsf{T}.$$

The minimizers of F are exactly the consensus vectors $x = xJ$. On the row-disagreement subspace $\{e : eJ = 0\}$, the matrix $I - P$ has eigenvalues $1 - \lambda_i(P)$, $i = 2, \dots, n$, and hence F is ρ-strongly convex and 2-smooth with respect to the Euclidean norm.

Algorithm. Applying Nesterov's accelerated gradient method to the quadratic disagreement objective $F(x) = \frac{1}{2}x(I-P)x^\mathsf{T}$ on the row-disagreement subspace gives the following algorithm.

Algorithm 3 Nesterov-accelerated gossip

1: **input:** Symmetric doubly-stochastic P, initial vector ξ, stepsize $\eta > 0$, and momentum $\beta \ge 0$.
2: Initialize $x_{-1} = x_0 = \xi$.
3: **for** $t = 0, 1, 2, \dots$ **do**
4: $\quad y_t = x_t + \beta(x_t - x_{t-1})$.
5: $\quad x_{t+1} = y_t - \eta y_t(I-P) = (1-\eta)y_t + \eta y_t P$.
6: **end for**

Equivalently, each node performs

$$y_{i,t} = x_{i,t} + \beta(x_{i,t} - x_{i,t-1}), \qquad x_{i,t+1} = (1-\eta)y_{i,t} + \eta \sum_j P_{ji} y_{j,t}.$$

Since P is symmetric, this is the usual weighted neighbor average.

Unlike classical Chebyshev gossip, which explicitly constructs the minimax degree-t polynomial on the disagreement spectrum, the recursion above arises from Nesterov acceleration on the quadratic $F(x) = \frac{1}{2}x(I-P)x^\top$. The resulting polynomial need not have Chebyshev-optimal constants, but it gives the same $O(\rho^{-1/2}\log(1/\varepsilon))$ communication dependence.

Convergence analysis. By directly utilizing standard convergence bounds for Nesterov's method in the strongly convex and smooth case, one can obtain the following.

Theorem 4 (Accelerated gossip convergence)**.** *Let P be symmetric and doubly stochastic, and suppose that the non-consensus eigenvalues of $I-P$ lie in $[\rho, 2]$ for some $\rho \in (0,1]$. Run Algorithm 3 with*

$$\eta = \frac{1}{2}, \qquad \beta = \frac{1-\sqrt{\rho/2}}{1+\sqrt{\rho/2}}.$$

Then, for every $t \geq 0$ it holds that $x_t J = \xi J$, and

$$\|x_t - \xi J\|_2 \leq \frac{2}{\sqrt{\rho}}\left(1-\sqrt{\rho/2}\right)^{t/2}\|\xi - \xi J\|_2.$$

Consequently, Algorithm 3 reaches $\|x_t - \xi J\|_2 \leq \varepsilon$ with iteration complexity

$$O\left(\rho^{-1/2}\log\left(\frac{\|\xi - \xi J\|_2}{\varepsilon\sqrt{\rho}}\right)\right).$$

We note that Ye et al. [2023, Proposition 1] provide a similar accelerated-consensus guarantee for a different recursion, which could serve as an alternative gossip primitive in our framework after adapting the gossip updates. Applied to the lazy matrix $W = (I+P)/2$, their bound offers a slightly different tradeoff between the prefactor and exponential contraction.

Proof. First, the average is invariant. Since P is doubly stochastic, $(I-P)J = 0$. Therefore $x_{t+1}J = y_t J$. Also,

$$y_t J = x_t J + \beta(x_t J - x_{t-1}J).$$

Since $x_{-1} = x_0 = \xi$, induction gives $x_t J = \xi J$ for all t.

Define the disagreement vector $e_t := x_t - \xi J$. The previous paragraph shows that $e_t J = 0$ for all t. Moreover, since $J(I-P) = 0$, we have $\xi J(I-P) = 0$. Subtracting ξJ from the update gives

$$e_{t+1} = \widetilde{y}_t - \frac{1}{2}\widetilde{y}_t(I-P), \qquad \widetilde{y}_t = e_t + \beta(e_t - e_{t-1}).$$

Thus the disagreement dynamics are exactly Nesterov's accelerated gradient method applied to the quadratic $F(e) := \frac{1}{2}e(I-P)e^\top$ on the row-disagreement subspace.

On this subspace, the eigenvalues of $I-P$ lie in $[\rho, 2]$. Hence F is ρ-strongly convex and 2-smooth with respect to the Euclidean norm. By the standard finite-time guarantee for Nesterov's accelerated gradient method on a μ-strongly convex and L-smooth objective, with $\eta = 1/L$ and $\beta = (1-\sqrt{\mu/L})/(1+\sqrt{\mu/L})$,

$$F(e_t) - F(e^\star) \leq \frac{L+\mu}{2}\left(1-\sqrt{\frac{\mu}{L}}\right)^t \|e_0 - e^\star\|_2^2.$$

Applying this bound with $\mu = \rho$, $L = 2$, $e^\star = 0$, and $e_0 = \xi - \xi J$, we obtain

$$F(e_t) \leq 2\left(1-\sqrt{\rho/2}\right)^t \|\xi - \xi J\|_2^2.$$

On the other hand, strong convexity on the row-disagreement subspace gives $F(e_t) \geq \frac{\rho}{2}\|e_t\|_2^2$. Combining the two inequalities yields

$$\|x_t - \xi J\|_2 \leq \frac{2}{\sqrt{\rho}}\left(1-\sqrt{\rho/2}\right)^{t/2}\|\xi - \xi J\|_2.$$

The iteration-complexity statement follows by taking logarithms and using $(1-a)^k \leq \exp(-ak)$ for $a \in (0,1)$. □

D Proofs for Decentralized Doubly Accelerated SGD (DDA-SGD)

For the analysis, define the network averages

$$\bar{x}_t := \frac{1}{M} X_t \mathbf{1}, \qquad \bar{z}_t := \frac{1}{M} Z_t \mathbf{1}, \qquad \bar{q}_t := \frac{1}{M} Q_t \mathbf{1},$$

and the averaged stochastic oracle

$$\widehat{G}_t := \frac{1}{M} \sum_{i=1}^{M} g_{t,i}.$$

We write

$$\zeta_t := \widehat{G}_t - \nabla f(\bar{q}_t), \qquad b_t := \mathbb{E}[\zeta_t \mid \mathcal{F}_t].$$

Before proving the main convergence result, we give some preliminary lemmas. First we give a bound on the heterogeneity of the local gradients away from the optimum, which is used to control the bias of the averaged oracle induced by delayed gossip.

Lemma 5 (Heterogeneity away from the optimum)**.** *Under Eq.* (1)*, every square-integrable random vector X satisfies*

$$\left(\mathbb{E}\left[\frac{1}{M} \sum_{i=1}^{M} \|\nabla f_i(X) - \nabla f(X)\|^2 \right] \right)^{1/2} \le \zeta_\star + 2L \left(\mathbb{E}\left\|X - x^\star\right\|^2 \right)^{1/2}.$$

Proof. For any deterministic $x \in \mathbb{R}^d$,

$$\begin{aligned} &\left(\frac{1}{M} \sum_{i=1}^{M} \|\nabla f_i(x) - \nabla f(x)\|^2 \right)^{1/2} \\ &\quad \le \left(\frac{1}{M} \sum_{i=1}^{M} \left\|\nabla f_i(x^\star) - \nabla f(x^\star)\right\|^2 \right)^{1/2} \\ &\qquad + \left(\frac{1}{M} \sum_{i=1}^{M} \left\|\nabla f_i(x) - \nabla f_i(x^\star)\right\|^2 \right)^{1/2} + \left\|\nabla f(x) - \nabla f(x^\star)\right\|. \end{aligned}$$

The first term is at most $\zeta_\star$ by Eq. (1); the last two terms are at most $L\|x - x^\star\|$ by L-smoothness of the f_i's and of f. Therefore the deterministic heterogeneity at x is at most $\zeta_\star + 2L\|x - x^\star\|$. Applying this pointwise with $x = X$ and using the triangle inequality in L_2 gives the claim. □

Next we show that the network averages satisfy the same delayed accelerated recursions as the single-node algorithm (Algorithm 2), with a single stochastic oracle given by the average of the local stochastic gradients. We also give bounds on the bias and variance of this averaged oracle in terms of the network disagreement.

Lemma 6 (Average oracle induced by delayed gossip)**.** *The network averages satisfy*

$$\bar{q}_{t+1} = \bar{x}_t + m_t(1 + m_{t+1})(\bar{x}_t - \bar{x}_{t-1}), \qquad \bar{z}_{t+1} = \bar{z}_t - \eta_t \widehat{G}_t,$$
$$\bar{x}_{t+1} = (1 - \theta_t)\bar{x}_t + \theta_t \bar{z}_{t+1}.$$

In particular, when $\eta_t = a_t/\beta$ for all t, these are the same delayed accelerated recursions as Algorithm 2, with stochastic oracle $\widehat{G}_t$ and query point $\bar{q}_t$. Moreover,

$$\|b_t\| \le \frac{L}{\sqrt{M}} \left\|Q_t - \bar{Q}_t\right\|_F, \qquad \mathbb{E}[\|\zeta_t - b_t\|^2 \mid \mathcal{F}_t] \le \frac{\sigma^2}{MB}. \tag{18}$$

Proof. Since $PJ = J$, each accelerated gossip block preserves network averages. Thus

$$\bar{q}_{t+1} = \bar{x}_t + m_t(1 + m_{t+1})(\bar{x}_t - \bar{x}_{t-1}).$$

Also, by averaging the local updates,

$$\bar{z}_{t+1} = \bar{z}_t - \eta_t \widehat{G}_t, \qquad \bar{x}_{t+1} = (1 - \theta_t)\bar{x}_t + \theta_t \bar{z}_{t+1}.$$

When $\eta_t = a_t/\beta$ for all t, these are precisely the single-node delayed accelerated recursions for the averaged variables.

The conditional bias is

$$b_t = \frac{1}{M}\sum_{i=1}^{M}\left(\nabla f_i(q_{t,i}) - \nabla f_i(\bar{q}_t)\right).$$

By L-smoothness and Cauchy-Schwarz inequality,

$$\|b_t\| \le \frac{L}{M}\sum_{i=1}^{M}\left\|q_{t,i} - \bar{q}_t\right\| \le \frac{L}{\sqrt{M}}\left\|Q_t - \bar{Q}_t\right\|_F.$$

The centered stochastic part is the average of MB conditionally independent zero-mean samples, each with conditional second moment at most σ^2, hence its conditional second moment is at most $\sigma^2/(MB)$. □

Finally, we give a bound on the network bias and output disagreement induced by delayed gossip. The bound depends on the initial disagreement, the heterogeneity of the local gradients away from the optimum, and the pathwise bound on the network average.

Lemma 7 (RMS network bias and output disagreement from delayed gossip)**.** *Assume Eq.* (1)*. Suppose the algorithmic parameters obey, for* $0 \le t < T$*,*

$$\eta_t \ge 0, \qquad 0 \le \theta_t \le 1, \qquad 0 \le m_t(1 + m_{t+1}) \le 2,$$

and that each accelerated gossip block uses

$$\alpha_{\mathrm{g}} = \frac{1}{2}, \qquad \beta_{\mathrm{g}} = \frac{1 - \sqrt{\rho/2}}{1 + \sqrt{\rho/2}}.$$

Let

$$Q_T^{\mathrm{rms}} := \max_{0\le t<T}\left(\mathbb{E}\left\|\bar{q}_t - x^\star\right\|^2\right)^{1/2}, \qquad S_T^\eta := \sum_{i=0}^{T-1}\eta_i, \qquad \tau := \frac{2}{\sqrt{\rho}}\left(1 - \sqrt{\rho/2}\right)^{B/2}.$$

Define the initial disagreement quantities

$$D_{\mathrm{init}}^{\mathrm{rms}} := \left(\mathbb{E}\left\|X_0 - \bar{X}_0\right\|_F^2\right)^{1/2} + \left(\mathbb{E}\left\|Z_0 - \bar{Z}_0\right\|_F^2\right)^{1/2},$$

$$Q_{\mathrm{init}}^{\mathrm{rms}} := \max\left\{\left(\mathbb{E}\left\|Q_0 - \bar{Q}_0\right\|_F^2\right)^{1/2}, \left(\mathbb{E}\left\|\widetilde{X}_{-1} - \bar{\widetilde{X}}_{-1}\right\|_F^2\right)^{1/2}\right\}, \qquad \Delta_{\mathrm{init}}^{\mathrm{rms}} := D_{\mathrm{init}}^{\mathrm{rms}} + 3LS_T^\eta Q_{\mathrm{init}}^{\mathrm{rms}}.$$

If

$$5LS_T^\eta\tau \le \frac{1}{2}, \tag{19}$$

then, for every $0 \le t < T$*,*

$$\left(\mathbb{E}\,\|b_t\|^2\right)^{1/2} \le 10LS_T^\eta\tau\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right) + \frac{10L\tau\Delta_{\mathrm{init}}^{\mathrm{rms}} + 3LQ_{\mathrm{init}}^{\mathrm{rms}}}{\sqrt{M}}. \tag{20}$$

Moreover, the final gossiped output satisfies

$$\left(\mathbb{E}\left\|\widetilde{X}_T - \bar{X}_T\right\|_F^2\right)^{1/2} \le 2S_T^\eta\tau\sqrt{M}\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right) + 2\tau\Delta_{\mathrm{init}}^{\mathrm{rms}}. \tag{21}$$

Proof. Define

$$D_T^{\mathrm{rms}} := \max_{0\le t\le T}\left(\mathbb{E}\left\|X_t - \bar{X}_t\right\|_F^2\right)^{1/2}.$$

By Eq. (2), the non-consensus eigenvalues of $I - P$ lie in $[\rho, 2]$. Thus Theorem 4 applies directly. Applying it row by row to each accelerated gossip block gives preservation of averages and the Frobenius contraction factor τ. Let $J := \mathbf{1}\mathbf{1}^\top/M$ and write $W^\perp := W(I - J) = W - \bar{W}$ for the

disagreement component of any matrix $W \in \mathbb{R}^{d\times M}$. First we bound the local gradient disagreement. Since $G_t^{\perp}$ is the distance from G_t to the consensus subspace,

$$\left\|G_t^{\perp}\right\|_F \le \left\|G_t - \nabla f(\bar{q}_t)\mathbf{1}^{\mathsf{T}}\right\|_F.$$

Writing $g_{t,i} = B^{-1}\sum_{s=1}^{B} G_i(q_{t,i}, \xi_{t,i}^{(s)})$, the ith column of the matrix on the right is

$$g_{t,i} - \nabla f(\bar{q}_t) = \big(g_{t,i} - \nabla f_i(q_{t,i})\big) + \big(\nabla f_i(q_{t,i}) - \nabla f_i(\bar{q}_t)\big) + \big(\nabla f_i(\bar{q}_t) - \nabla f(\bar{q}_t)\big).$$

For $t \ge 1$, since

$$Q_t = \widetilde{X}_{t-1} + m_{t-1}(1+m_t)(\widetilde{X}_{t-1} - \widetilde{X}_{t-2})$$

and $0 \le m_{t-1}(1+m_t) \le 2$, the accelerated-gossip contraction gives

$$\left(\mathbb{E}\left\|Q_t - \bar{Q}_t\right\|_F^2\right)^{1/2} \le 5\tau D_T^{\mathrm{rms}} + 2Q_{\mathrm{init}}^{\mathrm{rms}}.$$

The same bound, with the right-hand side enlarged to $5\tau D_T^{\mathrm{rms}} + 3Q_{\mathrm{init}}^{\mathrm{rms}}$, also holds at $t = 0$. By the triangle inequality in L_2, the minibatch variance bound, L-smoothness, the preceding display, and Lemma 5,

$$\left(\mathbb{E}\left\|G_t^{\perp}\right\|_F^2\right)^{1/2} \le \sqrt{M}\,\frac{\sigma}{\sqrt{B}} + 5L\tau D_T^{\mathrm{rms}} + 3LQ_{\mathrm{init}}^{\mathrm{rms}} + \sqrt{M}\left(\zeta_\star + 2LQ_T^{\mathrm{rms}}\right). \tag{22}$$

We now express primal disagreement through past gradient disagreements:

$$Z_{t+1}^{\perp} = Z_t^{\perp} - \eta_t G_t^{\perp}, \qquad X_{t+1}^{\perp} = (1-\theta_t)X_t^{\perp} + \theta_t Z_{t+1}^{\perp}.$$

Because $0 \le \theta_t \le 1$,

$$\left\|X_t^{\perp}\right\|_F \le \left\|X_0^{\perp}\right\|_F + \left\|Z_0^{\perp}\right\|_F + \sum_{i=0}^{t-1}\eta_i\left\|G_i^{\perp}\right\|_F, \qquad 0 \le t \le T,$$

pathwise. Taking L_2 norms, using the triangle inequality (in L_2), and applying Eq. (22) gives

$$D_T^{\mathrm{rms}} \le D_{\mathrm{init}}^{\mathrm{rms}} + S_T^{\eta}\left(5L\tau D_T^{\mathrm{rms}} + 3LQ_{\mathrm{init}}^{\mathrm{rms}} + \sqrt{M}\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right)\right).$$

Under Eq. (19), moving the self-coupling term to the left gives

$$D_T^{\mathrm{rms}} \le 2\Delta_{\mathrm{init}}^{\mathrm{rms}} + 2S_T^{\eta}\sqrt{M}\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right).$$

Since $\left\|\widetilde{X}_T - \bar{X}_T\right\|_F \le \tau\left\|X_T - \bar{X}_T\right\|_F$, this proves Eq. (21). Finally, for every $0 \le t < T$, by Eq. (18) and the query-disagreement bound above,

$$\|b_t\| \le \frac{L}{\sqrt{M}}\left\|Q_t - \bar{Q}_t\right\|_F \le \frac{5L\tau}{\sqrt{M}}D_T^{\mathrm{rms}} + \frac{3LQ_{\mathrm{init}}^{\mathrm{rms}}}{\sqrt{M}},$$

and using the preceding bound on D_T^{rms} proves Eq. (20). □

We proceed to prove the main convergence result for Algorithm 1.

D.1 Proof of Theorem 1

Proof. **Stage 1: Reduction to a single-node problem.** By Lemma 6, the averaged dynamics satisfy the single-node delayed accelerated recursion analyzed in Theorem 3. Namely, with averaged oracle $\widehat{G}_t = \nabla f(\bar{q}_t) + \zeta_t$, the sequence $(\bar{q}_t, \bar{z}_t, \bar{x}_t)$ satisfies, for $0 \le t < T$,

$$\bar{q}_{t+1} = \bar{x}_t + m_t(1+m_{t+1})(\bar{x}_t - \bar{x}_{t-1}), \qquad \bar{z}_{t+1} = \bar{z}_t - \eta_t\widehat{G}_t,$$
$$\bar{x}_{t+1} = (1-\theta_t)\bar{x}_t + \theta_t\bar{z}_{t+1}.$$

The same lemma gives the centered-variance part of Eq. (4), namely

$$\mathbb{E}[\|\zeta_t - b_t\|^2 \mid \mathcal{F}_t] \le \frac{\sigma^2}{MB}.$$

Since the chosen parameters have $\eta_t = a_t/\beta$, and since the consensual initialization gives $\bar{x}_{-1} = \bar{x}_0 = \bar{z}_0 = \bar{q}_0 = x_0$, the assumption $R \ge \|x_0 - x^\star\|$ gives the radius condition $R \ge \|\bar{x}_0 - x^\star\|$. Thus the averaged dynamics have the initialization and recursion required by Theorem 3.

Stage 2: Verifying the hypotheses of Theorem 3. In order to invoke Theorem 3, it remains to verify the following two properties:

(i) The averaged oracle calls are square-integrable, as required in Eq. (5);

(ii) The averaged oracle admits deterministic RMS bias bounds $(\epsilon_t)_{t=0}^{T-1}$, completing Eq. (4);

We first prove property (i). The optimality condition gives $\nabla f(x^\star) = 0$, so Eq. (1) bounds the average squared norm of the vectors $\nabla f_i(x^\star)$. Together with L-smoothness and the local variance assumption, this implies by induction that all local iterates, query points, and averaged oracle calls generated up to time T are square-integrable.

We next prove property (ii). The goal is to show that the RMS oracle bias $(\mathbb{E}\,\|b_t\|^2)^{1/2}$ is uniformly bounded for all $0 \le t < T$. Let

$$\tau := \frac{2}{\sqrt{\rho}}\left(1 - \sqrt{\rho/2}\right)^{B/2}, \qquad S_T^\eta := \sum_{t=0}^{T-1} \eta_t, \qquad Q_T^{\mathrm{rms}} := \max_{0 \le t < T} \left(\mathbb{E}\left\|\bar{q}_t - x^\star\right\|^2\right)^{1/2}.$$

The chosen parameters satisfy $\eta_t \ge 0$, $0 \le \theta_t \le 1$, and $0 \le m_t(1 + m_{t+1}) \le 2$ for $0 \le t < T$, and the gossip parameters are those required by Lemma 7. The initialization in Algorithm 1 is consensual, so $D_{\mathrm{init}}^{\mathrm{rms}} = Q_{\mathrm{init}}^{\mathrm{rms}} = \Delta_{\mathrm{init}}^{\mathrm{rms}} = 0$ in Lemma 7. Also, Q_0 is consensual, so $b_0 = 0$ by Eq. (18). Thus Lemma 7 will apply once we check that $5LS_T^\eta\tau \le 1/2$.

To check this inequality, put

$$U_N := \frac{N+M}{M}, \qquad \lambda_N := \frac{32N}{\sqrt{M}}\left(1 + \frac{\zeta_\star + \sigma}{LR}\right).$$

By definition, $\Lambda_N \ge e\lambda_N$. Since $B \ge 1$ and $T \ge 1$, we have $N \ge M$, $\lambda_N \ge 1$, $U_N \le 2N/M$, $T = N/(MB) \le N/M$, and $T + 1 \le U_N$. Hence

$$10LA_{T-1} = \frac{5}{128}T(T+1) \le \left(\frac{N+M}{M}\right)^2 \le \lambda_N^7.$$

By the definition of τ, the inequality $1 - x \le \exp(-x)$ for $x \in [0,1]$, and the choice of B,

$$\tau \le \frac{2}{\sqrt{\rho}}\exp\left(-\frac{\sqrt{\rho}}{2\sqrt{2}}B\right) \le \frac{2}{\sqrt{\rho}}\left(\frac{\rho}{e\lambda_N}\right)^7 \le \lambda_N^{-7}.$$

Since $S_T^\eta = A_{T-1}/\beta$ and $\beta \ge 1$, this implies

$$5LS_T^\eta\tau \le 5LA_{T-1}\tau \le \frac{1}{2}.$$

Therefore Lemma 7 applies and gives

$$\left(\mathbb{E}\,\|b_t\|^2\right)^{1/2} \le 10LS_T^\eta\tau\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right) \le \bar{\epsilon},$$

for every $0 \le t < T$, where

$$\bar{\epsilon} := 10LA_{T-1}\tau(\zeta_\star + \sigma + 2LQ_T^{\mathrm{rms}}).$$

We next close the radius dependence in $\bar{\epsilon}$. Define the crude no-bias radius

$$C_T := 2^{10}(T+1)\left(R + \sqrt{\frac{\sigma R}{L\sqrt{MB}}}\,T^{3/4}\right).$$

Applying Lemma 4 to the averaged dynamics with oracle variance parameter $\sigma^2/(MB)$ and uniform RMS bias bound $\bar{\epsilon}$ gives

$$Q_T^{\mathrm{rms}} \le C_T + 2^{10}(T+1)^2 A_{T-1}\bar{\epsilon}.$$

Moreover,

$$T^2(T+1)^4 \le \left(\frac{N+N}{M}\right)^6 \le \lambda_N^7,$$

and hence

$$c_T := 20 \cdot 2^{10} L^2 A_{T-1}^2 (T+1)^2 \tau = \frac{5}{16} T^2 (T+1)^4 \tau \le \frac{5}{16}.$$

Substituting the definition of $\bar{\epsilon}$ into the preceding radius bound gives

$$\begin{aligned} Q_T^{\mathrm{rms}} &\le C_T + 2^{10}(T+1)^2 A_{T-1}\bar{\epsilon} \\ &= C_T + 2^{10}(T+1)^2 A_{T-1} \cdot 10 L A_{T-1} \tau (\zeta_\star + \sigma + 2L Q_T^{\mathrm{rms}}) \\ &= C_T + 10 \cdot 2^{10} L A_{T-1}^2 (T+1)^2 \tau (\zeta_\star + \sigma) + c_T Q_T^{\mathrm{rms}}. \end{aligned}$$

Since $c_T \le 1/2$,

$$Q_T^{\mathrm{rms}} \le 2C_T + 20 \cdot 2^{10} L A_{T-1}^2 (T+1)^2 \tau (\zeta_\star + \sigma).$$

Substituting the resulting bound on Q_T^{rms} back into the expression $\zeta_\star + \sigma + 2LQ_T^{\mathrm{rms}}$ and using $2c_T \le 5/8 < 3$ shows that this quantity is at most a constant multiple of its bias-free analogue:

$$\begin{aligned} \zeta_\star + \sigma + 2LQ_T^{\mathrm{rms}} &\le \zeta_\star + \sigma + 4LC_T + 40 \cdot 2^{10} L^2 A_{T-1}^2 (T+1)^2 \tau (\zeta_\star + \sigma) \\ &= (1 + 2c_T)(\zeta_\star + \sigma) + 4LC_T \\ &\le 4(\zeta_\star + \sigma) + 4LC_T = 4(\zeta_\star + \sigma + LC_T). \end{aligned}$$

Consequently,

$$\begin{aligned} \bar{\epsilon} &= 10 L A_{T-1} \tau (\zeta_\star + \sigma + 2L Q_T^{\mathrm{rms}}) \\ &\le 10 L A_{T-1} \tau \cdot 4(\zeta_\star + \sigma + LC_T) \\ &= 40 L A_{T-1} \tau (\zeta_\star + \sigma + LC_T). \end{aligned}$$

It remains to make this last bound explicit. Using $B \ge 1$, $T \le U_N$, and $T + 1 \le U_N$,

$$C_T \le 2^{10} U_N \left(R + \sqrt{\frac{\sigma R}{L}}\, U_N^{3/4} \right).$$

Since

$$\sqrt{\frac{\sigma}{LR}} \le 1 + \frac{\zeta_\star + \sigma}{LR},$$

and $T^4(T+1) \le U_N^5$, we get

$$\begin{aligned} \zeta_\star + \sigma + LC_T &\le LR \left(\frac{\zeta_\star + \sigma}{LR} + 2^{10} U_N + 2^{10} U_N^{7/4} \sqrt{\frac{\sigma}{LR}} \right) \\ &\le \frac{LR}{2^{11} T^4 (T+1)} 2^{11} U_N^5 \left(\frac{\zeta_\star + \sigma}{LR} + 2^{10} U_N + 2^{10} U_N^{7/4} \sqrt{\frac{\sigma}{LR}} \right) \\ &\le \frac{LR}{2^{11} T^4 (T+1)} \left(2^{16} + 2^{27} + 2^{28} \right) \left(\frac{N}{M} \right)^7 \left(1 + \frac{\zeta_\star + \sigma}{LR} \right)^7 \\ &\le \frac{LR\lambda_N^7}{2^{11}\sqrt{M}\, T^4 (T+1)}. \end{aligned}$$

Combining this estimate with $\tau \le \lambda_N^{-7}$ and $A_{T-1} = T(T+1)/(256L)$ gives

$$\begin{aligned} \bar{\epsilon} &\le 40 L A_{T-1} \tau (\zeta_\star + \sigma + LC_T) \\ &\le 40L \cdot \frac{T(T+1)}{256L} \cdot \lambda_N^{-7} \cdot \frac{LR\lambda_N^7}{2^{11}\sqrt{M}\, T^4 (T+1)} \\ &= \frac{40}{2^{19}} \frac{LR}{\sqrt{M}\, T^3} \le 2^{-13} \frac{LR}{\sqrt{M}\, T^3}. \end{aligned}$$

Thus property (ii) holds with deterministic RMS bias bounds

$$\epsilon_0 := 0, \qquad \epsilon_t := \bar{\epsilon}, \qquad 1 \le t < T.$$

Stage 3: Applying Theorem 3 and absorbing the network error. By the recursion, radius condition, centered variance bound, and properties (i)–(ii), the hypotheses of Theorem 3 hold for the averaged dynamics with oracle variance parameter $\sigma^2/(MB)$ and the deterministic RMS bias bounds above. Let

$$\Gamma_T^{\mathrm{net}} := \frac{1}{A_{T-1}} \sum_{t=0}^{T-1} a_t \epsilon_t.$$

Applying Theorem 3 to the averaged dynamics and substituting $T = N/(MB)$ gives

$$\mathbb{E}[f(\bar{x}_T) - f(x^\star)] \le \frac{6\sigma R}{\sqrt{N}} + \frac{128 L R^2 M^2 B^2}{N^2} + \mathcal{E}_N^{\mathrm{net}},$$

where

$$\mathcal{E}_N^{\mathrm{net}} := 2^{11} \left(R + \sqrt{\frac{\sigma R}{L}} \frac{N^{3/4}}{MB} \right) \frac{N}{MB} \Gamma_T^{\mathrm{net}} + \frac{64}{L} \left(\frac{N}{MB} \right)^4 (\Gamma_T^{\mathrm{net}})^2.$$

Absorbing the bias terms. Since $\epsilon_t \le 2^{-13} LR/T^3$ for every $0 \le t < T$, the weighted average satisfies

$$\Gamma_T^{\mathrm{net}} \le 2^{-13} \frac{LR}{T^3}.$$

We verify the absorption term by term. Let

$$S_N := \frac{\sigma R}{\sqrt{N}}, \qquad D_N := \frac{LR^2}{T^2} = \frac{LR^2 M^2 B^2}{N^2}.$$

If $\Gamma_T^{\mathrm{net}} \le 2^{-13} LR/T^3$, then

$$2^{11} RT \Gamma_T^{\mathrm{net}} \le \frac{1}{4} D_N, \qquad \frac{64}{L} T^4 (\Gamma_T^{\mathrm{net}})^2 \le 2^{-20} D_N.$$

For the mixed stochastic-bias term, using

$$\sqrt{\frac{\sigma R}{L}} \frac{N^{3/4}}{MB} = \sqrt{\frac{\sigma R}{L\sqrt{MB}}}\, T^{3/4}$$

and the same bound on Γ_T^{net} gives

$$2^{11} \sqrt{\frac{\sigma R}{L}} \frac{N^{3/4}}{MB} \cdot T \Gamma_T^{\mathrm{net}} \le \frac{1}{4} \sqrt{D_N S_N} \le \frac{1}{8} (D_N + S_N).$$

Combining these estimates gives $\mathcal{E}_N^{\mathrm{net}} \le D_N + S_N$. Substituting this into the preliminary bound gives the averaged-iterate estimate

$$\mathbb{E}[f(\bar{x}_T) - f(x^\star)] \le \frac{7\sigma R}{\sqrt{N}} + \frac{129 L R^2 M^2 B^2}{N^2}.$$

Stage 4: From the averaged iterate to local outputs. It remains to transfer the averaged-iterate guarantee to the local gossiped outputs. By Eq. (21) and the definition of $\bar{\epsilon}$,

$$\left(\mathbb{E} \left\| \widetilde{X}_T - \bar{X}_T \right\|_F^2 \right)^{1/2} \le \frac{\sqrt{M}}{5L} \bar{\epsilon},$$

and hence, for every node i,

$$\left(\mathbb{E} \left\| \widetilde{x}_{T,i} - \bar{x}_T \right\|^2 \right)^{1/2} \le \frac{\sqrt{M}}{5L} \bar{\epsilon} \le 2^{-13} \frac{R}{T^3}.$$

The preceding output-disagreement bound implies

$$L\mathbb{E}\left\|\widetilde{x}_{T,i}-\bar{x}_T\right\|^2 \le D_N.$$

For a fixed node i, the standard smooth convex inequality

$$f(y)-f(x^\star) \le 2(f(x)-f(x^\star))+L\|y-x\|^2$$

with $x=\bar{x}_T$ and $y=\widetilde{x}_{T,i}$ gives

$$\mathbb{E}[f(\widetilde{x}_{T,i})-f(x^\star)] \le \frac{14\sigma R}{\sqrt{N}}+259D_N.$$

Finally, since $N \ge M$ gives $\Lambda_N \ge 100$ and $\rho \le 1$,

$$B \le \frac{20}{\sqrt{\rho}}\log(\Lambda_N/\rho)+1 \le \frac{21}{\sqrt{\rho}}\log(\Lambda_N/\rho).$$

Thus

$$D_N \le \frac{441LR^2M^2}{\rho N^2}\log^2(\Lambda_N/\rho),$$

which gives the claimed bound since $259 \cdot 441 = 114219 < 2^{17}$. □

E The strongly convex case through restarting

This appendix records the restart reduction needed for strongly convex objectives. The only new ingredient is that, after the first epoch, the local copies used to initialize the next epoch need not be exactly consensual.

Algorithm 4 Restarted delayed accelerated gossip for strongly convex objectives

1: **Input:** common x_0, number of epochs K, epoch budgets $(N_s)_{s=0}^{K-1}$, epoch radii $(R_s)_{s=0}^{K-1}$, and the parameter rules of Theorem 1.
2: Set the initial matrix $U^{(0)}=x_0\mathbf{1}^\mathsf{T}$.
3: **for** $s=0,1,\dots,K-1$ **do**
4: Run one warm-started epoch from $X_{-1}=X_0=Z_0=Q_0=\widetilde{X}_{-1}=U^{(s)}$, using budget N_s and radius parameter R_s.
5: Let $U^{(s+1)}$ be the matrix of local outputs $\widetilde{X}^{(s)}$ produced by this epoch.
6: **end for**
7: Each node outputs its column of $U^{(K)}$.

Theorem 5 (Restarted rate for strongly convex objectives)**.** *Assume, in addition to the assumptions of Theorem 1, that f is μ-strongly convex. Let $\Delta_0 \ge f(x_0)-f(x^\star)$, fix $0<\epsilon\le\Delta_0$, and set*

$$K:=\left\lceil \log_2\frac{\Delta_0}{\epsilon}\right\rceil, \qquad \Delta_s:=\frac{\Delta_0}{2^s}, \qquad R_s:=\sqrt{\frac{2\Delta_0}{\mu\,2^s}}, \qquad 0\le s\le K.$$

Let

$$\ell_\epsilon:=\max\left\{60,\ \log\left[\frac{2^{100}}{\rho\sqrt{M}}\left(1+\frac{\zeta_\star+\sigma}{L\sqrt{\epsilon/\mu}}\right)\left(1+\frac{\sigma^2}{\mu\epsilon}+M\sqrt{\frac{L}{\mu\rho}}\right)\right]\right\},$$

and choose the epoch budget

$$N_s:=\left\lceil 2^{16}\frac{\sigma^2}{\mu\epsilon}+2^{18}M\sqrt{\frac{L}{\mu\rho}}\,\ell_\epsilon^2\right\rceil, \qquad 0\le s<K. \tag{23}$$

For these budgets define

$$\Lambda_s:=\frac{100N_s}{\sqrt{M}}\left(1+\frac{\zeta_\star+\sigma}{LR_s}\right), \qquad B_s:=\left\lceil\max\left\{1,\frac{20}{\sqrt{\rho}}\log\frac{\Lambda_s}{\rho}\right\}\right\rceil.$$

Assume for simplicity that $T_s := N_s/(MB_s)$ is a positive integer for every $0 \le s < K$. Run Algorithm 4 with these epoch budgets. Then the returned local outputs satisfy $\max_{1\le i\le M} \mathbb{E}[f(U_{\cdot i}^{(K)}) - f(x^\star)] \le \epsilon$ with a total sample complexity of

$$N = \sum_{s=0}^{K-1} N_s \le K\left(1 + 2^{16}\frac{\sigma^2}{\mu\epsilon} + 2^{18}M\sqrt{\frac{L}{\mu\rho}}\,\ell_\epsilon^2\right) = \widetilde{O}\left(\frac{\sigma^2}{\mu\epsilon} + M\sqrt{\frac{L}{\mu\rho}}\log\frac{\Delta_0}{\epsilon}\right).$$

In order to prove Theorem 5, we need the following warm-started version of Theorem 1. We defer its proof until the end of this section, after the proof of Theorem 5.

Theorem 6 (Warm-started distributed epoch)**.** *Assume the setting of Section 2. Let $U \in \mathbb{R}^{d\times M}$ be an initialization matrix and write $\bar{U} := U\mathbf{1}\mathbf{1}^\mathsf{T}/M$. Assume $R^2 \ge \mathbb{E}\left\|\bar{U}_{\cdot,1} - x^\star\right\|^2$, define*

$$\Lambda_N := \frac{100N}{\sqrt{M}}\left(1 + \frac{\zeta_\star + \sigma}{LR}\right),$$

and choose $B := \left\lceil \max\left\{1,\ (20/\sqrt{\rho})\log(\Lambda_N/\rho)\right\}\right\rceil$. Assume for simplicity that $T := N/(MB)$ is a positive integer, and let η_t be the stepsize prescribed by Theorem 1 for this choice of R, N, B, T. Let $S_T^\eta := \sum_{t=0}^{T-1}\eta_t$, $\tau := \frac{2}{\sqrt{\rho}}\left(1 - \sqrt{\rho/2}\right)^{B/2}$, and assume that the RMS column disagreement of the initialization satisfies

$$\left(\mathbb{E}\left\|U - \bar{U}\right\|_F^2\right)^{1/2} \le 2^{-16}\frac{R}{T}, \qquad \tau(2 + 3LS_T^\eta)\left(\mathbb{E}\left\|U - \bar{U}\right\|_F^2\right)^{1/2} \le 2^{-24}\frac{R}{T^3}. \tag{24}$$

Run Algorithm 1 with the same parameter choices as in Theorem 1, except that the initialization line is replaced by

$$X_{-1} = X_0 = Z_0 = Q_0 = \widetilde{X}_{-1} = U.$$

Then, for every node i,

$$\mathbb{E}[f(\widetilde{x}_{T,i}) - f(x^\star)] \le \frac{20\sigma R}{\sqrt{N}} + \frac{2^{21}LR^2M^2}{\rho N^2}\log^2\frac{\Lambda_N}{\rho}.$$

Moreover, the final output disagreement satisfies

$$\left(\mathbb{E}\left\|\widetilde{X}_T - \bar{X}_T\right\|_F^2\right)^{1/2} \le 2^{-11}\frac{R}{T^3}. \tag{25}$$

We are ready to prove the convergence for the restarted method.

Proof of Theorem 5. For the chosen budgets, let

$$\tau_s := \frac{2}{\sqrt{\rho}}\left(1 - \sqrt{\rho/2}\right)^{B_s/2}, \qquad S_s^\eta := \sum_{t=0}^{T_s-1}\eta_t,$$

where η_t is the epoch stepsize used with radius R_s. Set $d_0 := 0$, and for $1 \le s < K$ put

$$d_s := 2^{-11}\frac{R_{s-1}}{T_{s-1}^3}, \qquad 1 \le s < K.$$

We now check that the explicit budget is large enough for the two restart requirements: the inherited disagreement is small enough for the next warm-started epoch, and the current epoch reduces the error by a factor two. For $0 \le s < K$, $R_s \ge \sqrt{\epsilon/\mu}$, and Eq. (23) gives

$$N_s \le 2^{19}\ell_\epsilon^2\left(1 + \frac{\sigma^2}{\mu\epsilon} + M\sqrt{\frac{L}{\mu\rho}}\right).$$

By the definition of ℓ_ϵ and the slack factor 2^{100},

$$\log\frac{\Lambda_s}{\rho} \le 2\ell_\epsilon.$$

Indeed, after taking logarithms, the only extra contribution beyond the quantity inside ℓ_ϵ is $2\log\ell_\epsilon + \log(100\cdot 2^{19})$, which is at most ℓ_ϵ because $\ell_\epsilon \geq 60$. Consequently

$$B_s \leq \frac{41}{\sqrt{\rho}}\ell_\epsilon, \qquad T_s = \frac{N_s}{MB_s} \geq \frac{2^{18}}{41}\sqrt{\frac{L}{\mu}}\ell_\epsilon \geq 16.$$

Here we used the standard consequence $\mu \leq L$ of μ-strong convexity and L-smoothness. Since N_s is independent of s and R_s is decreasing, B_s is nondecreasing and therefore T_s is nonincreasing.

The inherited disagreement then satisfies

$$d_s \leq 2^{-16}\frac{R_s}{T_s}, \qquad \tau_s(2+3LS_s^\eta)d_s \leq 2^{-24}\frac{R_s}{T_s^3}, \tag{26}$$

for every $1 \leq s < K$. Indeed, $R_{s-1} = \sqrt{2}R_s$ and $T_{s-1} \geq T_s$, so the first inequality follows from $T_s \geq 16$. For the second, put

$$\lambda_s := \frac{32N_s}{\sqrt{M}}\left(1 + \frac{\zeta_\star + \sigma}{LR_s}\right).$$

As in the proof of Theorem 1, $\Lambda_s \geq e\lambda_s$ and the choice of B_s imply $\tau_s \leq \lambda_s^{-7}$, while $T_s^2(T_s+1)^4 \leq \lambda_s^7$. Since $LS_s^\eta \leq LA_{T_s-1} = T_s(T_s+1)/256$,

$$\tau_s(2+3LS_s^\eta) \leq \frac{4}{(T_s+1)^4} \leq 2^{-14},$$

and the second inequality in Eq. (26) follows from the first display defining d_s.

The same explicit budget also gives the epoch-accuracy bound

$$\frac{20\sigma R_s}{\sqrt{N_s}} + \frac{2^{21}LR_s^2M^2}{\rho N_s^2}\log^2\frac{\Lambda_s}{\rho} \leq \frac{\Delta_0}{2^{s+1}}, \tag{27}$$

for every $0 \leq s < K$. The stochastic term is at most $(20\sqrt{2}/2^8)\Delta_s \leq \Delta_s/4$, because $\Delta_s \geq \epsilon$. The network term is at most

$$\frac{2^{23}LR_s^2M^2}{\rho N_s^2}\ell_\epsilon^2 \leq 2^{-12}\frac{\Delta_s}{\ell_\epsilon^2} \leq \frac{\Delta_s}{4}.$$

We prove by induction that, at the beginning of epoch s,

$$\mathbb{E}[f(\bar{U}_{\cdot 1}^{(s)}) - f(x^\star)] \leq \Delta_s, \qquad \mathbb{E}\left\|\bar{U}_{\cdot 1}^{(s)} - x^\star\right\|^2 \leq R_s^2, \qquad \left(\mathbb{E}\left\|U^{(s)} - \bar{U}^{(s)}\right\|_F^2\right)^{1/2} \leq d_s.$$

The claim is true at $s = 0$ by the definition of Δ_0 and strong convexity, and by $d_0 = 0$. Suppose it holds at epoch s. The inherited disagreement bound together with Eq. (26) verifies the warm-start smallness assumption of Theorem 6 for the initialization $U^{(s)}$. Applying that theorem with radius R_s, and then using Eq. (27), gives, for every node i,

$$\mathbb{E}[f(U_{\cdot i}^{(s+1)}) - f(x^\star)] \leq \Delta_{s+1}.$$

By convexity, the network average at the next restart also satisfies

$$\mathbb{E}[f(\bar{U}_{\cdot 1}^{(s+1)}) - f(x^\star)] \leq \frac{1}{M}\sum_{i=1}^M \mathbb{E}[f(U_{\cdot i}^{(s+1)}) - f(x^\star)] \leq \Delta_{s+1}.$$

Strong convexity then gives

$$\mathbb{E}\left\|\bar{U}_{\cdot 1}^{(s+1)} - x^\star\right\|^2 \leq \frac{2\Delta_{s+1}}{\mu} = R_{s+1}^2.$$

If $s < K-1$, then the output-disagreement estimate Eq. (25) gives

$$\left(\mathbb{E}\left\|U^{(s+1)} - \bar{U}^{(s+1)}\right\|_F^2\right)^{1/2} \leq 2^{-11}\frac{R_s}{T_s^3} = d_{s+1},$$

which closes the induction.

After K epochs, $\Delta_K \leq \epsilon$, proving the accuracy claim. The sample-complexity bound is the explicit sum of the constant epoch budgets in Eq. (23); the factors K and ℓ_ϵ^2 are logarithmic in the target accuracy and problem parameters, giving the stated $\widetilde{O}$ form. □

E.1 Proof of Theorem 6

Let

$$D_U^{\mathrm{rms}} := \left(\mathbb{E}\left\|U - \bar{U}\right\|_F^2\right)^{1/2}.$$

For the warm initialization above,

$$D_{\mathrm{init}}^{\mathrm{rms}} = 2D_U^{\mathrm{rms}}, \qquad Q_{\mathrm{init}}^{\mathrm{rms}} = D_U^{\mathrm{rms}}, \qquad \Delta_{\mathrm{init}}^{\mathrm{rms}} = (2 + 3LS_T^{\eta})D_U^{\mathrm{rms}}.$$

The direct effect of the initial query occurs only at $t = 0$. Indeed,

$$\left(\mathbb{E}\left\|b_0\right\|^2\right)^{1/2} \le \frac{L}{\sqrt{M}} D_U^{\mathrm{rms}},$$

whereas for $t \ge 1$ the query is formed from already-gossiped primal copies. Writing $D_T^{\mathrm{rms}} := \max_{0\le t\le T}(\mathbb{E}\left\|X_t - \bar{X}_t\right\|_F^2)^{1/2}$, the accelerated-gossip contraction gives

$$\left(\mathbb{E}\left\|Q_t - \bar{Q}_t\right\|_F^2\right)^{1/2} \le 5\tau D_T^{\mathrm{rms}}, \qquad 1 \le t < T.$$

Repeating the proof of Lemma 7 with this sharper query-disagreement bound, and using the same verification of $5LS_T^{\eta}\tau \le 1/2$ as in the proof of Theorem 1, gives, for $t \ge 1$,

$$\left(\mathbb{E}\left\|b_t\right\|^2\right)^{1/2} \le 10LS_T^{\eta}\tau\left(\zeta_\star + \frac{\sigma}{\sqrt{B}} + 2LQ_T^{\mathrm{rms}}\right) + \frac{10L\tau\Delta_{\mathrm{init}}^{\mathrm{rms}}}{\sqrt{M}}.$$

The assumption Eq. (24) implies

$$\frac{a_0}{A_{T-1}}\frac{L}{\sqrt{M}} D_U^{\mathrm{rms}} \le 2^{-15}\frac{LR}{T^3}, \qquad \frac{10L\tau\Delta_{\mathrm{init}}^{\mathrm{rms}}}{\sqrt{M}} \le 2^{-20}\frac{LR}{T^3}, \qquad 2\tau\Delta_{\mathrm{init}}^{\mathrm{rms}} \le 2^{-23}\frac{R}{T^3}.$$

Thus the weighted RMS bias budget in Theorem 3 is enlarged by at most $2^{-14}LR/T^3$ compared with the consensual-start proof of Theorem 1. Repeating the same radius-closing and bias-absorption argument with this larger bias budget gives

$$\mathbb{E}[f(\bar{x}_T) - f(x^\star)] \le \frac{10\sigma R}{\sqrt{N}} + \frac{2^{20}LR^2M^2}{\rho N^2}\log^2\frac{\Lambda_N}{\rho}.$$

Here we use the RMS-start version of Theorem 3, which follows from the same proof by taking expectations in the initial term $\frac{\beta}{2}\left\|x^\star - x_0\right\|^2$; hence it is enough that $R^2 \ge \mathbb{E}\left\|\bar{U}_{\cdot,1} - x^\star\right\|^2$.

The output-disagreement part of Lemma 7 gives the usual term from the proof of Theorem 1, plus an additive term of at most $2^{-23}R/T^3$. Hence

$$\left(\mathbb{E}\left\|\widetilde{X}_T - \bar{X}_T\right\|_F^2\right)^{1/2} \le 2^{-11}\frac{R}{T^3},$$

after enlarging constants. The same smoothness comparison between $\widetilde{x}_{T,i}$ and $\bar{x}_T$ used at the end of Theorem 1's proof then yields the stated local bound. □

F A Lower Bound for Linear-Span Decentralized Methods

This appendix proves the lower bound stated in Theorem 2 for linear-span decentralized methods. The linear-span restriction is important—it rules out methods that guess arbitrary directions unrelated to the gradients and messages seen so far—but it is a standard setting in which zero-chain lower-bound constructions are established [e.g., Nesterov, 2018].

Consider the path graph on M nodes, with $M \ge 6$. Let $P_{\mathrm{path}} = I - \frac{1}{4}\mathsf{L}_{\mathrm{path}}$, where $\mathsf{L}_{\mathrm{path}}$ is the unnormalized Laplacian of the path. This is a symmetric doubly stochastic gossip matrix supported on the path, and its spectral gap satisfies, by the standard spectrum of the path Laplacian [Chung, 1997],

$$\rho = \frac{1}{2}\left(1 - \cos\frac{\pi}{M}\right) = \Theta(M^{-2}).$$

A linear-span decentralized first-order method is allowed, in each communication round, to perform arbitrary local linear algebra on vectors already held by a node, query the local gradient at points in the span of those vectors, and send arbitrary vectors from this span to neighboring nodes. Equivalently, if $\mathcal{S}_i(t)$ is the subspace available to node i after t gossip communication rounds, then $\mathcal{S}_i(0) = \{0\}$, and $\mathcal{S}_i(t+1)$ is contained in the span of $\mathcal{S}_i(t)$, the neighboring subspaces $\mathcal{S}_j(t)$, $\{i,j\} \in E$, and local gradients evaluated at points in that span. All nodes are initialized at the origin. The proof below gives the method the additional power of free local span closure inside each block, so the lower bound also applies to the one-query-per-round linear-span model.

Proof of Theorem 2. Let $m := \lfloor M/3 \rfloor$, and partition the path into

$$V_L := \{1, \dots, m\}, \qquad V_R := \{M - m + 1, \dots, M\},$$

and the remaining middle block. The graph distance between V_L and V_R is

$$\Delta := M - 2m + 1 = \Theta(M) = \Theta(\rho^{-1/2}).$$

Set

$$s := 2 + \left\lfloor \frac{S}{\Delta} \right\rfloor, \qquad d = 2(s+1), \qquad a := \frac{R}{\sqrt{d}}.$$

For $z \in \mathbb{R}^d$, define

$$H_L(z) := z_1^2 - 2z_1 + \sum_{\substack{1 \le r < d \\ r \text{ odd}}} (z_r - z_{r+1})^2, \qquad H_R(z) := \sum_{\substack{1 \le r < d \\ r \text{ even}}} (z_r - z_{r+1})^2,$$

and $H := H_L + H_R$. Fix a sufficiently small universal constant $c_0 > 0$, for instance $c_0 = 2^{-6}$, and define the local objectives by

$$f_i(x) := c_0 L \frac{M}{m} a^2 H_L(x/a), \qquad i \in V_L,$$

$$f_i(x) := c_0 L \frac{M}{m} a^2 H_R(x/a), \qquad i \in V_R,$$

and $f_i(x) := 0$ for all middle-block nodes. Since $m \ge M/4$ for $M \ge 6$, and the Hessians of H_L and H_R have operator norm bounded by an absolute constant, the choice of c_0 ensures that each f_i is convex and L-smooth. Averaging over the nodes gives

$$f(x) = c_0 L a^2 H(x/a).$$

The vector $z^\star = \mathbf{1} \in \mathbb{R}^d$ minimizes both H_L and H_R, and hence $x^\star := a\mathbf{1}$ minimizes f. Moreover, $\|x^\star\| = a\sqrt{d} = R$.

We next record the zero-chain property. Let

$$E_r := \operatorname{span}\{e_1, \dots, e_r\}, \qquad E_0 := \{0\}.$$

The odd-even split of the coordinate couplings is what turns the usual zero-chain construction into a communication lower bound. If $z \in E_r$, then a gradient of H_L can introduce the new coordinate e_{r+1} only when r is odd, while a gradient of H_R can introduce the new coordinate e_{r+1} only when r is even. This follows directly from the fact that H_L contains only the odd couplings $(z_r - z_{r+1})^2$, whereas H_R contains only the even couplings. The left objective contains the linear term and the couplings $(1,2), (3,4), \dots$, while the right objective contains the couplings $(2,3), (4,5), \dots$. Thus the left block can start the chain and reveal coordinate 2, but the next coupling needed to reveal coordinate 3 is available only in the right block. After the right block reveals coordinate 3, the next coupling needed to reveal coordinate 4 is again available only in the left block, and so on.

For $r \ge 2$, coordinate $r+1$ must be generated in V_R when r is even and in V_L when r is odd, from a vector whose r-th coordinate is already nonzero. These generating blocks alternate. Since a message can move by at most one edge per communication round, information needs at least Δ rounds to travel from one generating block to the other. Therefore, even in the strengthened model that allows arbitrary local span closure inside each block, no coordinate beyond E_s can appear anywhere in the network after S gossip communication rounds. Hence every local output, and also their average $\bar{x}_S$, belongs to E_s.

It remains to lower bound the objective gap of vectors in E_s. For $z = x/a$, the identity

$$H(z) - H(\mathbf{1}) = (z_1 - 1)^2 + \sum_{r=1}^{d-1} (z_r - z_{r+1})^2$$

holds. If $x \in E_s$, then $z \in E_s$, so $z_{s+1} = 0$. By Cauchy-Schwarz inequality,

$$1 = (1 - z_1) + (z_1 - z_2) + \cdots + (z_s - z_{s+1})$$

implies

$$(z_1 - 1)^2 + \sum_{r=1}^{s} (z_r - z_{r+1})^2 \geq \frac{1}{s+1}.$$

Consequently, every $x \in E_s$ satisfies

$$f(x) - f(x^\star) = c_0 L a^2 \big(H(x/a) - H(\mathbf{1})\big) \geq \frac{c_0 L R^2}{d(s+1)} = \frac{c_0 L R^2}{2(s+1)^2}.$$

Since $s + 1 \leq 3 + S/\Delta$ and $\Delta = \Theta(\rho^{-1/2})$, the last display is bounded below by

$$c L R^2 \min\left\{1, \frac{1}{\rho S^2}\right\}$$

for another universal constant $c > 0$. Applying this to $x = \bar{x}_S \in E_s$ proves the theorem. □

G Numerical experiments

In this section, we provide the full details of our experimental evaluation (Section 5). We follow the synthetic least-squares benchmark of Eisen et al. [2025]. The experiments address two questions. First, how does DDA-SGD compare with D-SGD and DAT-SGD [Eisen et al., 2025] under the same sample and communication budget? Second, how sensitive is DDA-SGD to the block length B? This is the only parameter of Algorithm 1 without a counterpart in previous methods, and it simultaneously determines the local minibatch size, the number of accelerated gossip rounds per macro step and, through $T = N/(MB) - 1$, the number of updates. All results are averaged over three random seeds.

Problem. The local objective of worker i is $f_i(x) = \frac{1}{2}\|A_i x - b_i\|^2$, where $A_i \in \mathbb{R}^{d\times d}$ has i.i.d. standard normal entries and $b_i = A_i(x^\sharp - \delta_i)$. The vector $x^\sharp \sim \mathcal{N}(0, \frac{1}{d} I_d)$ is sampled once per configuration, and $\delta_i \sim \mathcal{N}(0, \frac{\zeta^2}{d} I_d)$ introduces heterogeneity across workers. A stochastic gradient query at x returns $\nabla f_i(x) + \xi$ with $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{d} I_d)$. We set $d = 50$, $\sigma^2, \zeta^2 \in \{1, 10\}$ and $M \in \{4, 9, 25, 49, 100\}$. All workers start from a common point $x_0 \sim \mathcal{N}(0, I_d)$, as Algorithm 1 requires, so the initial error is about 50. Following Eisen et al. [2025], we report the mean squared distance $\frac{1}{M}\sum_{i=1}^{M} \|x_i - x^\star\|^2$ between the local outputs x_i and the minimizer $x^\star$ of f.

Networks. We use the ring and the torus of Eisen et al. [2025]. On the ring, each worker averages itself and its two neighbors with weight $\frac{1}{3}$; on the two-dimensional torus, itself and its four neighbors with weight $\frac{1}{5}$ (for $M = 4$, the torus reduces to the ring). Both gossip matrices are fixed, symmetric and doubly stochastic; their spectral gaps are listed in Table 4.

Budget and methods. Every method runs for 10^5 communication rounds, and in each round every worker draws one stochastic gradient, so that $N = 10^5 M$. D-SGD returns its last iterate and DAT-SGD its averaged iterate, as in Eisen et al. [2025]; DAT-SGD exchanges two vectors per round, whereas D-SGD and DDA-SGD exchange one. With block length B, DDA-SGD performs $T = 10^5/B - 1$ updates: the final block of B gradients is charged to the budget but, since $\eta_T = \theta_T = 0$, does not affect the output $\widetilde{x}_{T,i}$. We run DDA-SGD with two parameter schedules, both normalized so that the effective primal step $\theta_t \eta_t$ equals the tuned stepsize η during updates.

- *Theorem schedule*: the momentum and averaging schedules of Theorem 1, with tuned stepsize and block length: $\alpha_{\mathrm{g}} = \frac{1}{2}$, $\beta_{\mathrm{g}} = (1 - \sqrt{\rho/2})/(1 + \sqrt{\rho/2})$, $\theta_t = 2/(t+2)$, $m_t = (t-1)/(t+2)$ and $\eta_t = \eta(t+2)/2$.
- *Constant momenta*: a practical variant with $\theta_t \equiv 0.1$, $m_t \equiv 0.9$ and $\eta_t \equiv \eta/0.1$, and with $\alpha_{\mathrm{g}} = \frac{1}{2}$, $\beta_{\mathrm{g}} = 0.9$ on every graph. The momenta satisfy the same relation $m_t = \theta_t(1 - \theta_{t-1})/\theta_{t-1}$ as in Theorem 1.

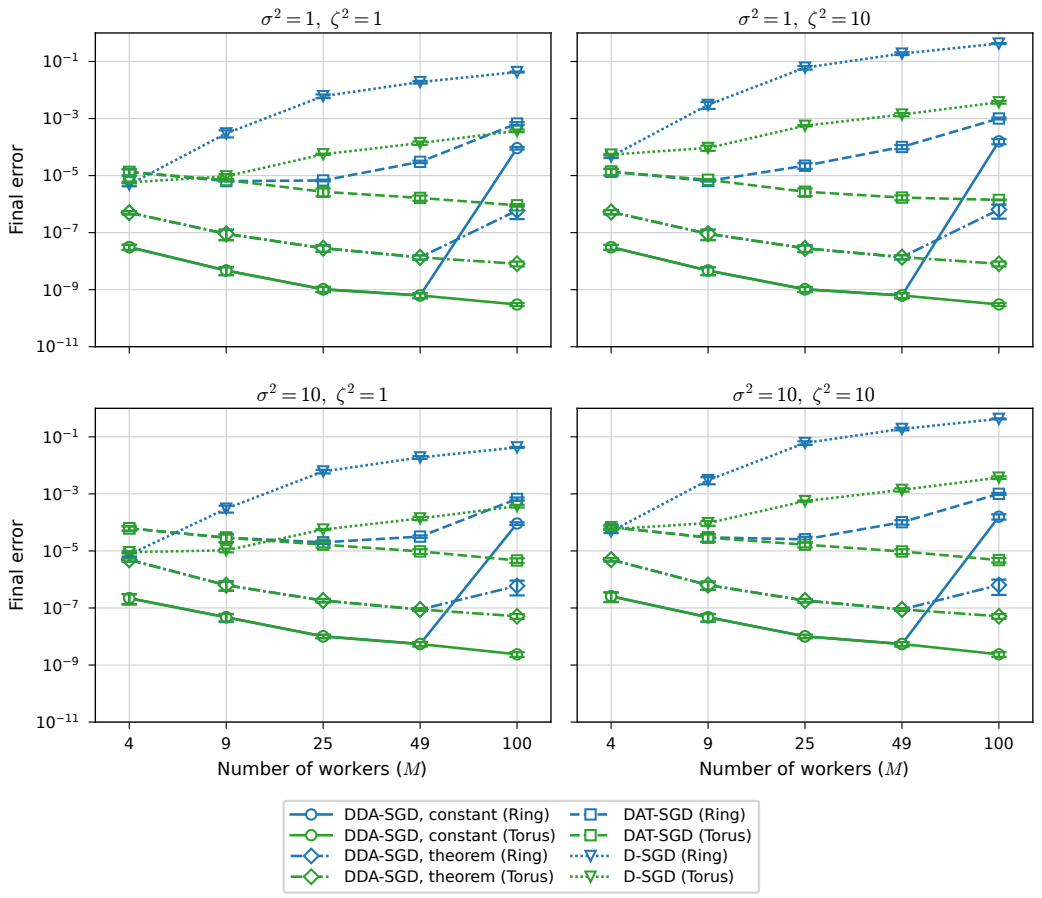

Figure 2: Final error against the number of workers for each pair $(\sigma^2, \zeta^2) \in \{1, 10\}^2$, in the layout of Eisen et al. [2025, Figure 1]. Color denotes the topology and line style the method. The plotted quantity is $\frac{1}{M}\sum_{i=1}^{M} \left\|x_i - x^\star\right\|^2$, with x_i the local output of the method. Error bars show the sample standard deviation across three random seeds. Table 3 lists the case $\sigma^2 = \zeta^2 = 1$.

Tuning. As in Eisen et al. [2025], for every method, topology, M, σ^2 and ζ^2, we select the stepsize $\eta \in \{10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}\}$ with the lowest final error. For DDA-SGD, we select B jointly with η from the twenty values

$$B \in \{1, 2, 4, 5, 10, 16, 20, 25, 50, 100, 200, 250, 400, 500, 625, 1000, 1250, 2000, 2500, 5000\}.$$

This gives DDA-SGD a larger tuning budget than the baselines. We select hyperparameters by minimizing mean final error over three random seeds and report the mean and sample standard deviation over those same seeds.

Compute. The experiments ran on one computer with 10 CPU cores and 32 GB of RAM. The block-length grid took approximately two hours of wall-clock time, and the comparison with both baselines took approximately 1.5 hours on four cores.

G.1 Comparison with D-SGD and DAT-SGD

Figure 2 shows both schedules of DDA-SGD, together with D-SGD and DAT-SGD, for every pair $(\sigma^2, \zeta^2) \in \{1, 10\}^2$, in the layout of Eisen et al. [2025, Figure 1]. Table 3 records the case $\sigma^2 = \zeta^2 = 1$, with the block length selected for DDA-SGD.

Optimization performance. Figure 2 shows that DDA-SGD achieves lower error than both baselines across all tested ring and torus configurations. This holds for both momentum schedules. With

Graph	M	D-SGD	DAT-SGD	DDA-SGD (constant)	DDA-SGD (theorem)
Ring	4	$4.8(0.6) \cdot 10^{-6}$	$1.3(0.3) \cdot 10^{-5}$	$\mathbf{3.1(0.6)} \cdot \mathbf{10^{-8}}$ (500)	$5.0(0.6) \cdot 10^{-7}$ (200)
	9	$3.0(0.8) \cdot 10^{-4}$	$6.4(0.2) \cdot 10^{-6}$	$\mathbf{4.7(1.4)} \cdot \mathbf{10^{-9}}$ (250)	$9.1(3.6) \cdot 10^{-8}$ (250)
	25	$6.1(0.9) \cdot 10^{-3}$	$6.8(0.5) \cdot 10^{-6}$	$\mathbf{1.0(0.2)} \cdot \mathbf{10^{-9}}$ (400)	$2.8(0.7) \cdot 10^{-8}$ (400)
	49	$1.9(0.2) \cdot 10^{-2}$	$3.0(0.3) \cdot 10^{-5}$	$\mathbf{6.3(1.3)} \cdot \mathbf{10^{-10}}$ (400)	$1.4(0.2) \cdot 10^{-8}$ (400)
	100	$4.3(0.2) \cdot 10^{-2}$	$6.7(0.8) \cdot 10^{-4}$	$9.2(1.0) \cdot 10^{-5}$ (1000)	$\mathbf{6.0(3.0)} \cdot \mathbf{10^{-7}}$ (625)
Torus	4	$5.8(0.4) \cdot 10^{-6}$	$1.3(0.3) \cdot 10^{-5}$	$\mathbf{3.1(0.6)} \cdot \mathbf{10^{-8}}$ (500)	$5.0(0.6) \cdot 10^{-7}$ (200)
	9	$9.4(1.8) \cdot 10^{-6}$	$7.0(0.6) \cdot 10^{-6}$	$\mathbf{4.7(1.4)} \cdot \mathbf{10^{-9}}$ (250)	$9.1(3.6) \cdot 10^{-8}$ (250)
	25	$5.6(0.3) \cdot 10^{-5}$	$2.7(0.8) \cdot 10^{-6}$	$\mathbf{1.0(0.2)} \cdot \mathbf{10^{-9}}$ (400)	$2.8(0.7) \cdot 10^{-8}$ (400)
	49	$1.4(0.2) \cdot 10^{-4}$	$1.7(0.3) \cdot 10^{-6}$	$\mathbf{6.3(1.3)} \cdot \mathbf{10^{-10}}$ (400)	$1.4(0.2) \cdot 10^{-8}$ (400)
	100	$3.7(0.4) \cdot 10^{-4}$	$9.1(0.9) \cdot 10^{-7}$	$\mathbf{3.0(0.4)} \cdot \mathbf{10^{-10}}$ (400)	$8.1(1.6) \cdot 10^{-9}$ (200)

Table 3: Final error for $\sigma^2 = \zeta^2 = 1$, the top-left panel of Fig. 2. The parenthetical number on each error is the standard deviation across three seeds, in the same units. The integer after a DDA-SGD entry is the selected block length. The lowest error in each row is in bold.

constant momenta, the final error is at least 6 times smaller than that of DAT-SGD and more than 30 times smaller than that of D-SGD. With the theorem-based schedule, the improvement over DAT-SGD is at least 12-fold. More precisely, constant momenta improve on DAT-SGD by a factor between 6 and $1.6 \cdot 10^5$ (median $1.6 \cdot 10^3$), and the theorem schedule improves on DAT-SGD by a factor between 12.7 and $7.3 \cdot 10^3$ (median 94).

Parallelism. At a fixed per-worker budget, DDA-SGD improves on the torus through $M = 100$ in all four panels of Fig. 2; at $\sigma^2 = \zeta^2 = 1$, error falls a hundredfold from $M = 4$. On the ring, it improves through $M = 49$ before error rises at $M = 100$, consistent with network error overtaking statistical error. The theorem-based schedule mitigates this rise, from $1.4 \cdot 10^{-8}$ to $6.0 \cdot 10^{-7}$ (Table 3). D-SGD deteriorates on both networks from the smallest M, especially at $\zeta^2 = 10$. DAT-SGD improves on the torus, while its lowest ring error occurs at $M = 9$ or $M = 25$, depending on the noise setting. Thus, DDA-SGD benefits from additional workers over the widest range on the ring.

Noise versus heterogeneity. The four panels separate the two error sources. Passing from the left column to the right multiplies ζ^2 by ten, and passing from the top row to the bottom multiplies σ^2 by ten. The curves of DDA-SGD with constant momenta barely move with ζ^2 (median ratio 1.00; the largest ratio, 1.74, occurs on the 100-worker ring) and rise by a median factor of 8.7 with σ^2. Thus, in most tested configurations, these trends are consistent with variance-dominated error and weak sensitivity to heterogeneity, as expected when the residual disagreement is negligible. D-SGD exhibits the opposite pattern (median ratios 1.0 in σ^2 and 10 in ζ^2), so its error is dominated by heterogeneity-induced disagreement, while DAT-SGD lies in between (median ratios 4.6 and 1.03).

Summary. On the ring and the torus, DDA-SGD is consistently and substantially more accurate than D-SGD and DAT-SGD, and benefits from additional workers over the widest range of M, provided that B is long enough for the gossip block to reach consensus, yet short enough to leave sufficiently many updates. The theorem schedule is robust and excels where communication is the bottleneck, whereas constant momenta are more accurate elsewhere.

G.2 Sensitivity to the block length

Figure 3 shows the final error of DDA-SGD as a function of B for both schedules, with η tuned separately for each B, and Table 4 lists the minimizing block lengths. We fix $\sigma^2 = \zeta^2 = 1$; the other variance levels shift the curves vertically but barely move their minimizers.

Three regimes. On the ring and the torus, the curves in Fig. 3 share one shape under both schedules. For $B \leq 5$, no configuration attains an error below one. On the 100-worker ring the error stays above one through $B = 50$ with constant momenta and through $B = 100$ with the theorem schedule. Once B is large the error rises again; $B = 5000$ leaves only $T = 19$ updates. The minimizing block lengths lie in $[200, 625]$, except on the 100-worker ring with constant momenta, where $B^\star = 1000$ (Table 4).

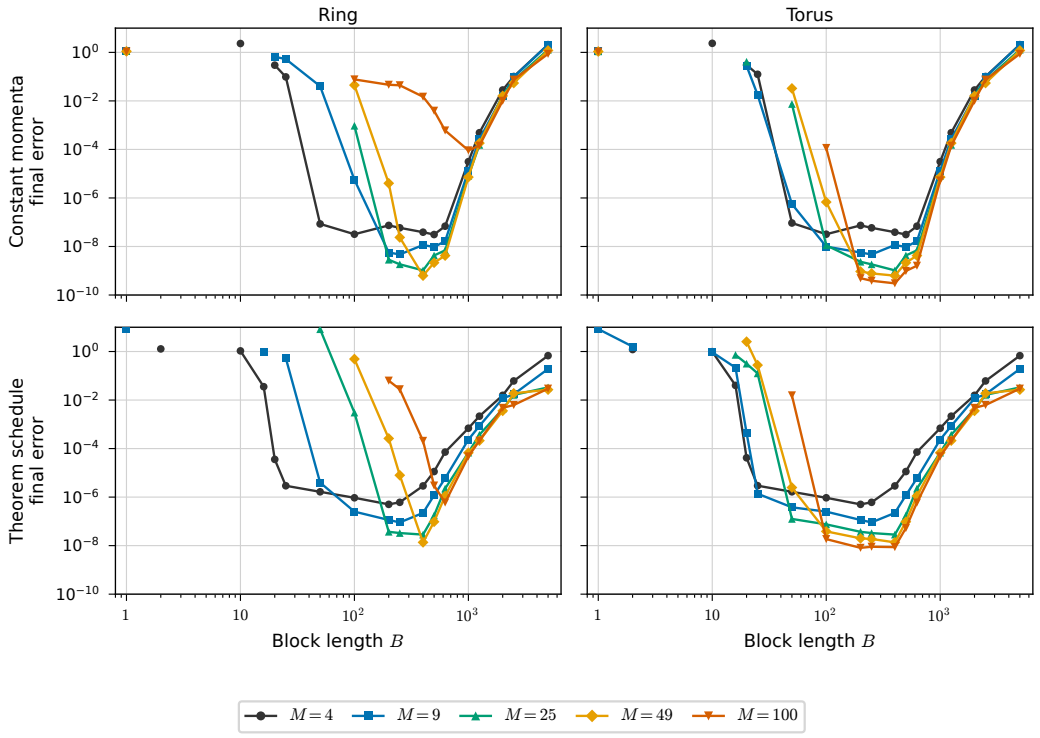

Figure 3: Final error of DDA-SGD against the block length B for $\sigma^2 = \zeta^2 = 1$. The top row uses constant momenta and the bottom row the schedule of Theorem 1; color denotes the number of workers. Both axes are logarithmic, and the stepsize is tuned separately at each B. Runs that diverged, or whose error exceeds 10, are omitted, which opens gaps in the curves.

				Constant momenta		Theorem schedule	
Graph	M	ρ	B_{thm}	$B^\star$	range	$B^\star$	range
Ring	4	0.67	422	500	50–625	200	25–400
	9	0.16	966	250	200–625	250	100–400
	25	0.021	2985	400	200–625	400	200–500
	49	0.0055	6292	400	400–625	400	400–500
	100	0.0013	13816	1000	625–1250	625	500–625
Torus	4	0.67	422	500	50–625	200	25–400
	9	0.60	458	250	100–625	250	50–400
	25	0.28	724	400	200–625	400	50–500
	49	0.15	1029	400	200–625	400	100–500
	100	0.076	1520	400	200–625	200	100–500

Table 4: Block-length selection for DDA-SGD ($\sigma^2 = \zeta^2 = 1$). ρ is the spectral gap. B_{thm} is the block length prescribed by Theorem 1, with the factor $1 + (\zeta_\star + \sigma)/(LR)$ in Λ_N set to one. $B^\star$ is the minimizing block length, and the range lists the block lengths of the grid whose error is within a factor of ten of the best.

Block lengths within a factor of ten of the best span a factor between 1.25 and 16 in B, about three in a typical configuration.

Consensus. At these block lengths, the ring and the torus produce nearly identical errors for every $M \le 49$ (Table 3), although their spectral gaps differ by a factor of up to 28. For $M \le 49$ and $\sigma^2 = \zeta^2 = 1$, the consensus distance $\frac{1}{M}\sum_{i=1}^{M} \|x_i - \bar{x}\|^2$ accounts for less than 10^{-4} of the final error, and DDA-SGD behaves as centralized accelerated minibatch SGD with minibatch size MB. On the ring, $B^\star$ grows from 250 at $M = 9$ ($\rho = 0.16$) to 1000 at $M = 100$ ($\rho = 1.3 \cdot 10^{-3}$); on the torus it remains between 200 and 500. The block length prescribed by Theorem 1 is comparable to the measured minimizers at $M = 4$, and exceeds them by more than an order of magnitude on the 100-worker ring.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction highlight the main contributions of the paper, including: (i) an accelerated decentralized method achieving $M \lesssim \sqrt{\rho}\, N^{3/4}$ critical parallelism, (ii) a matching lower bound for linear-span methods, and (iii) a strongly convex extension. The claims match the theoretical results established in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper is primarily theoretical. In Sections 1–2 we discuss the assumptions we make and their significance, and in the discussion and limitations paragraph we discuss the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide complete and correct proofs for all the results presented. Additionally, the assumptions used are provided in Section 2 or explicitly stated in the corresponding theorem.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup is described in detail in Section 5 and Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Section 5 and Appendix G detail the experimental setup: decentralized optimization on a simple synthetic least-squares problem adapted from Eisen et al. [2025]. The paper does not release code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings and hyperparameter selection procedure are described in Section 5 and Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix G.1 reports the full comparisons with D-SGD and DAT-SGD, with error bars indicating the sample standard deviation across three random seeds. Section 5 presents a compact version of these results.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix G reports the hardware and approximate wall-clock times for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper's contribution is primarily theoretical, and the research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This paper is foundational theoretical research on optimization algorithms. It is not tied to a specific application, and we do not foresee a direct path to negative societal impacts beyond those inherent to any general improvement in optimization methods.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any models, data, or other artifacts that pose a risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use any external code, data, or pre-trained models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets such as code, data, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.