Quantized Stochastic Primal–Dual Methods for Distributed Optimization under Relaxed Global Geometry

Susmit Sarkar[1,*] **Abhinav Raghuvanshi**[2,*] **Kushal Chakrabarti**[2] **Mayank Baranwal**[1,2]

[1]Indian Institute of Technology Bombay, Mumbai, Maharashtra, India
[2]Tata Consultancy Services Research, Thane, Maharashtra, India
[*]Equal Contribution

Abstract

We study distributed optimization with stochastic gradients and finite-bit communication modeled by random (unbiased) quantization. We propose `q−PDGD`, a quantized stochastic primal–dual method, and analyze it under relaxed global geometry. Under restricted secant inequality (RSI), a constant step-size yields linear contraction to an explicit neighborhood determined by gradient noise, quantization distortion, and network connectivity, while a diminishing step-size achieves $\mathcal{O}(1/k)$ convergence without shared-minimizer assumptions. Under Polyak–Łojasiewicz (PL) inequality, we obtain linear-to-neighborhood convergence in the same stochastic quantized setting. Our results match the best-known centralized stochastic rates in oracle complexity, and are supported by experiments demonstrating the predicted tradeoffs between quantization level, step-size choice, and graph structure.

1 INTRODUCTION

Distributed optimization is a central primitive in large-scale learning and networked control, enabling agents to cooperatively minimize a global objective while keeping data and computation local [Yang et al., 2019]. It underpins federated and data-parallel learning [Yu et al., 2024], large-scale empirical risk minimization [Jahani et al., 2020], distributed estimation in sensor/IoT networks [Chong et al., 2018], and network resource allocation/control [Nedić and Liu, 2018]. In many such systems, communication is the dominant bottleneck, motivating finite-bit (*quantized*) protocols.

Quantization has been studied extensively, with early control-oriented approaches using deterministic uniform quantizers (often with saturation) in consensus/ADMM/primal-dual message passing, which introduces bias and steady-state errors that must be controlled [Zhu et al., 2016]. Modern learning further requires *stochastic gradients* (mini-batches/noisy measurements), creating a second persistent error source. Recent results show that sharp sublinear rates can still be attainable under random quantization in distributed gradient methods with appropriate conditions and stepsizes; in particular, Dutta and Doan [2024] establishes $\mathcal{O}(1/k)$ convergence under specific assumptions. However, existing guarantees for stochastic quantized distributed optimization often emphasize server-based architectures or primal consensus/gradient-tracking schemes, and typically yield neighborhood convergence under constant step-sizes or rely on restrictive structural assumptions (e.g., shared minimizers) for exact convergence.

Motivated by these gaps, we study *primal-dual* methods for decentralized stochastic optimization under *random (unbiased) quantized* communication. We develop and analyze `q−PDGD`, providing convergence guarantees under relaxed global geometry (RSI/PL) and explicit tradeoffs among step-size, quantization resolution, network connectivity, and stochastic noise. Our main contributions are:

- **Stochastic quantized PDGD under RSI.** We establish convergence of a stochastic primal-dual gradient method with unbiased random quantization (`q−PDGD`) under RSI, without strong convexity or shared-minimizer assumptions.
- **Rates with constant and diminishing step-sizes.** For constant step-sizes, we prove linear contraction to an explicit noise/quantization/network-dependent neighborhood; for diminishing step-sizes, we obtain $\mathcal{O}(1/k)$ rates in the strongly-convex/RSI regime.
- **PL regime and experiments.** Under PL, we show linear-to-neighborhood convergence under stochasticity and quantization, and we empirically validate the predicted step-size/bit-rate tradeoffs and the linear-to-neighborhood vs. sublinear behavior.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5967–5996.

2 RELATED WORK

Finite-bit communication in distributed optimization has been studied from classic *deterministic* quantizers (often with clipping) to modern stochastic/structured compressors. Deterministic approaches appear in quantized consensus and ADMM-style primal–dual message passing, which characterize steady-state/consensus errors under finite-capacity links [Zhu et al., 2016]. Beyond fixed quantizers, algorithm-quantizer co-design yields stronger rates: Pu et al. [2016] develop iteratively-refining schemes for distributed proximal-gradient-type methods and show linear decay of quantization error and linear convergence under suitable interval conditions, while Rikos et al. [2023] combine gradient descent with a finite-time quantized-consensus primitive over directed graphs and obtain linear-to-neighborhood convergence for strongly convex objectives.

A complementary abstraction models quantization as a *compression operator* with either unbiasedness and bounded second moment or contractive/relative-error properties, unifying `QSGD` [Alistarh et al., 2017], `TernGrad` [Wen et al., 2017], and `signSGD` [Bernstein et al., 2018]. To mitigate bias and stabilize learning, *error-feedback* methods such as `EF21` provide clean guarantees, including improved rates for smooth nonconvex objectives and linear rates under PL-type conditions [Richtárik et al., 2021]. Related ideas compress gradient differences (`DIANA`) for sharper guarantees [Mishchenko et al., 2025] and use double quantization to reduce communication [Yu et al., 2019].

Rates also depend on architecture: server-based methods often admit explicit accuracy–communication tradeoffs under unbiased/difference compression [Alistarh et al., 2017, Mishchenko et al., 2025, Yu et al., 2019], whereas decentralized methods must control consensus/tracking error in addition to compression noise; `CHOCO-SGD` analyzes decentralized learning under broad compressor models and general nonconvex objectives [Koloskova* et al., 2020]. For constrained network optimization, quantized primal–dual dynamics with adaptive encoding/decoding can achieve linear convergence to the exact optimum for convex global costs while relating bandwidth to rate [Chen et al., 2023]. Finally, recent stochastic-approximation analyses of random quantization identify conditions/stepsizes achieving sharp sublinear rates, including $\mathcal{O}(1/k)$ [Dutta and Doan, 2024].

In contrast, we develop and analyze `q-PDGD` under RSI/PL with stochastic gradients and unbiased quantization, yielding explicit neighborhood and rate tradeoffs.

3 NOTATIONS & PRELIMINARIES

For any positive integer n, let $[n]$ denote the set $\{1, \dots, n\}$. We denote the concatenated column vector of vectors $z_i \in \mathbb{R}^{p_i}$ for $i \in [k]$ by $\text{col}(z_1, \dots, z_k)$. Let $\mathbf{1}_n$ and $\mathbf{0}_n$ denote the column vectors of all ones and all zeros of dimension n, respectively. $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Given a vector $[x_1, \dots, x_n]^\top \in \mathbb{R}^n$, $\text{diag}([x_1, \dots, x_n])$ represents a diagonal matrix with the i-th diagonal element being x_i. The notation $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

For a matrix $\boldsymbol{A}$, $\text{rank}(\boldsymbol{A})$, $\text{image}(\boldsymbol{A})$, and $\text{null}(\boldsymbol{A})$ denote its rank, image space, and null space, respectively. For two symmetric matrices $\boldsymbol{M}$ and $\boldsymbol{N}$, the inequality $\boldsymbol{M} \succeq \boldsymbol{N}$ indicates that $\boldsymbol{M} - \boldsymbol{N}$ is positive semi-definite. The spectral radius of a matrix is denoted by $\rho(\cdot)$, and $\rho_2(\cdot)$ indicates the smallest non-zero eigenvalue for matrices having positive eigenvalues. $\|\cdot\|$ represents the standard Euclidean norm for vectors or the induced 2-norm for matrices. For a given positive semi-definite matrix $\boldsymbol{A}$, the weighted norm is denoted by $\|\mathbf{x}\|_{\boldsymbol{A}} = \sqrt{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}$. For a differentiable function g, ∇g denotes the gradient of g. Finally, $\mathcal{P}_S(x) = \arg\min_{y \in S} \|x - y\|^2$ denotes the projection of a vector $x \in \mathbb{R}^n$ onto the set $S \subseteq \mathbb{R}^n$.

Definition 1. *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Restricted Secant Inequality (RSI) with constant $\nu > 0$ if $(\nabla f(x) - \nabla f(\mathcal{P}_{X^*}(x))^\top (x - \mathcal{P}_{X^*}(x)) \geq \nu \|x - \mathcal{P}_{X^*}(x)\|^2 \ \forall \mathbf{x} \in \mathbb{R}^n$, where X^* is the set of global minimizers of f [Shi et al., 2015, Yi et al., 2020a].*

Definition 2. *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Polyak–Łojasiewicz (PL) inequality with constant $\nu > 0$ if $\frac{1}{2}\|\nabla f(x)\|^2 \geq \nu\big(f(x) - f^*\big), \quad \forall x \in \mathbb{R}^n$, where $f^* = \min_{x \in \mathbb{R}^n} f(x)$ [Karimi et al., 2016, Yi et al., 2020b].*

The PL inequality is the weakest condition among others, such as strong convexity, essential strong convexity, weak strong convexity, and restricted secant inequality, that leads to *linear* convergence of deterministic centralized gradient-based methods [Karimi et al., 2016]. Both RSI and PL are weaker than strong convexity. For example, all linear least-squares problem $f(x) = \|Ax - b\|^2$ satisfy PL, but not necessarily strongly convex. Non-convex functions like $f(x) = x^2 + 3\sin^2(x)$, which satisfy PL, demonstrate the condition's generality beyond strong convexity. Similarly, all quasi-strongly convex functions or composition of strongly convex function with linear map plus a linear term satisfy RSI. One such example function is presented later in Section 6. Note that, neither of RSI and PL conditions require the function to be convex.

Definition 3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ be a weighted undirected graph, where $\mathcal{V} = [n]$ is the set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $A = A^\top = (a_{ij})$ is the weighted adjacency matrix with $a_{ij} \geq 0$ and $a_{ii} = 0$. Nodes i and j can communicate if $(i, j) \in \mathcal{E}$, i.e., $a_{ij} > 0$. The neighbor set and weighted degree of node i are defined as $\mathcal{N}_i = \{j \in [n] : a_{ij} > 0\}$, $\deg_i = \sum_{j=1}^n a_{ij}$. Let $D = \text{diag}([\deg_1, \dots, \deg_n])$ denote the degree matrix. The weighted Laplacian matrix is defined as $L = D - A$. The*

graph $\mathcal{G}$ is said to be connected if there exists a path between any pair of nodes.

4 PROBLEM SETUP AND THE q-PDGD ALGORITHM

Consider a network of n agents, where each agent $i \in \{1, \dots, n\}$ has a local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$. The agents collaborate to solve the following distributed optimization problem:

$$\min_{x \in \mathbb{R}^d} \; f(x) \coloneqq \sum_{i=1}^{n} f_i(x). \tag{1}$$

Let x^* denote an optimal solution of (1), and let $X^* \coloneqq \arg\min_{x \in \mathbb{R}^d} f(x)$ be the corresponding optimal set. Each agent maintains a local copy $x_i \in \mathbb{R}^d$ of the solution x^*. For simplicity, we define the stacked variables and function:

$$\mathbf{x} \coloneqq \mathrm{col}(x_1, \dots, x_n) \in \mathbb{R}^{nd}, \quad F(\mathbf{x}) \coloneqq \sum_{i=1}^{n} f_i(x_i). \tag{2}$$

The optimal set in the stacked space is $\mathcal{X}^* \coloneqq \{\mathbf{1}_n \otimes x^* : x^* \in X^*\}$. Further, we let $f^* \coloneqq \min_{x \in \mathbb{R}^d} f(x)$ denote the optimal value of the global objective.

Network Model. The communication among agents is described by an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ with $\mathcal{V} = \{1, \dots, n\}$ as defined by Definition 3. The graph is assumed to be connected. Let L denote the weighted Laplacian matrix associated with $\mathcal{G}$, and define the augmented Laplacian $\boldsymbol{L} \coloneqq L \otimes \boldsymbol{I}_d$.

Random Quantization Model. Each agent i quantizes its decision variable x_i using a quantization function $Q(\cdot)$, resulting in the quantized value $Q(x_i)$. We consider a quantization scheme characterized by the hyperparameters $\{l, u, b\}$, where $[l, u]$ denotes the quantization domain and b denotes the number of bits available for communication. We partition the interval $[l, u]$ into $B = 2^b - 1$ equal-length bins.[1] The length of each bin, referred to as the *quantization resolution*, is given by $\Delta \coloneqq \frac{u-l}{B}$. Let $\{\tau_m\}_{m=1}^{B+1}$ denote the quantization levels (endpoints), where $\tau_1 = l$ and $\tau_{B+1} = u$. Given a value $x \in [l, u]$, we identify the interval such that $x \in [\tau_i, \tau_{i+1})$. We define the relative location of x within the bin as $p = \frac{x - \tau_i}{\Delta}$. The stochastic quantizer $Q : [l, u] \to \{\tau_1, \dots, \tau_{B+1}\}$ is defined as:

$$Q(x) = \begin{cases} \tau_i, & \text{with probability } 1 - p, \\ \tau_{i+1}, & \text{with probability } p. \end{cases} \tag{3}$$

In this setting, at every iteration, agent i transmits the quantized value $Q(x_i)$, with its neighboring agents. For a vector $x \in \mathbb{R}^d$, quantization is applied element-wise. We refer to $\varepsilon(x) \coloneqq Q(x) - x$ as the *quantization error*.

[1] Note that $B = 2^b - 1$ bins result in $B + 1 = 2^b$ quantization levels, which can be uniquely represented by a b-bit codeword.

Lemma 1 ([Reisizadeh et al., 2018])**.** *For any input $z \in [l, u]$, the random quantization scheme $Q(\cdot)$ is unbiased and has bounded variance. Specifically, it satisfies the following properties:*

$$\mathbb{E}[Q(z) \mid z] = z, \qquad \mathbb{E}[|Q(z) - z|^2 \mid z] \le \frac{\Delta^2}{4}. \tag{4}$$

Quantization Noise. We define the quantization noise for agent i at iteration k as $\varepsilon_i(k) \coloneqq Q(x_i(k)) - x_i(k)$. We let $\mathcal{F}_k$ represent the filtration that contains the history of all variables generated by an algorithm up to iteration k. Since the quantization is performed element-wise on the vector $x_i(k) \in \mathbb{R}^d$, the properties established in Lemma 1 imply that the noise is zero-mean and has bounded variance conditioned on $\mathcal{F}_k$:

$$\mathbb{E}[\varepsilon_i(k) \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\varepsilon_i(k)\|^2 \mid \mathcal{F}_k] \le \frac{d\Delta^2}{4}. \tag{5}$$

4.1 ASSUMPTIONS

We adopt the following assumptions for the analysis.

Assumption 1. *The set of optimal solutions $X^* = \arg\min_{x \in \mathbb{R}^d} f(x)$ is nonempty and convex.*

Assumption 2. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and has a globally Lipschitz continuous gradient with constant $L_{f_i} > 0$, i.e., $\|\nabla f_i(a) - \nabla f_i(b)\| \le L_{f_i} \|a - b\|, \quad \forall a, b \in \mathbb{R}^d$.*

Assumption 3. *The global objective function $f(x)$ satisfies the Restricted Secant Inequality (RSI) condition with constant $\nu > 0$, i.e.,*

$$(\nabla f(x) - \nabla f(\mathcal{P}_{X^*}(x)))^\top (x - \mathcal{P}_{X^*}(x)) \ge \nu \|x - \mathcal{P}_{X^*}(x)\|^2, \quad \forall x \in \mathbb{R}^d. \tag{6}$$

Assumption 4. *The global objective function $f(x)$ satisfies the Polyak–Łojasiewicz (PL) inequality with constant $\nu > 0$, i.e.,*

$$\frac{1}{2}\|\nabla f(x)\|^2 \ge \nu\big(f(x) - f^*\big), \quad \forall x \in \mathbb{R}^d. \tag{7}$$

Assumption 5. *The set $\{\nabla F(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^*\}$ is a singleton.*

Assumption 6. *The communication graph $\mathcal{G}$ between agents is undirected and connected.*

Assumption 7. *Each agent has access to an unbiased stochastic gradient $g_i(x_i(k), \zeta_i(k)) = \nabla f_i(x_i(k)) + \zeta_i(k)$ with bounded variance:*

$$\mathbb{E}[\zeta_i(k) \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\zeta_i(k)\|^2 \mid \mathcal{F}_k] \le \sigma^2. \tag{8}$$

From Assumption 2, we define the maximum Lipschitz constant across all agents as $L_f := \max_{i\in[n]} L_{f_i}$. Assumption 5 implies that the aggregate gradient is invariant across the optimal set $\mathcal{X}^*$; specifically, the sum of local gradients at any optimal solution is unique. Finally, Assumption 6 ensures that the Laplacian L has a simple zero eigenvalue. Consequently, its second-smallest eigenvalue $\rho_2(L) > 0$.

Compared to the analysis of Distributed Consensus Stochastic Gradient (DCSG) method under the same quantization model from Dutta and Doan [2024], we do not assume X^* being singleton, and crucially, we do not assume that a global minimizer $x^* \in X^*$ also minimizes each f_i. Furthermore, the RSI condition in Assumption 3 and PL condition in Assumption 4 are both weaker than the strong convexity of global objective in Dutta and Doan [2024].

Now, we review below in Lemma 2 a prior result from Yi et al. [2020a], an implication of the RSI condition, that is pivotal for our key results presented in the next section.

Lemma 2 ([Yi et al., 2020a])**.** *Let Assumptions 1–3 hold. If* $\alpha > \frac{2nL_f^2+\nu L_f}{\nu\rho_2(L)}$*, then for all* $\mathbf{x} \in \mathbb{R}^{nd}$*,*

$$(\nabla F(\mathbf{x}) - \nabla F(\mathcal{P}_{\mathcal{X}^*}(\mathbf{x})))^\top(\mathbf{x} - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x})) + \alpha\mathbf{x}^\top \boldsymbol{L}\mathbf{x} \geq \nu_1 \left\|\mathbf{x} - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x})\right\|^2, \quad (9)$$

where $\nu_1 = \min\left\{\frac{\nu}{2n}, \alpha\rho_2(L) - \frac{2nL_f^2+\nu L_f}{\nu}\right\} > 0.$

4.2 STOCHASTIC QUANTIZED PDGD ALGORITHM

As discussed in Section 1, primal-dual gradient descent (PDGD) offers an effective framework for decentralized optimization. We now introduce q−PDGD, a quantized stochastic extension of PDGD whose deterministic roots trace back to [Kia et al., 2015, Yi et al., 2020a]. Our focus is the practically relevant regime with *simultaneous* inexactness from stochastic gradients and random (unbiased) quantized communication, as formalized in Section 4.

At each iteration $k \geq 0$, each agent i updates its primal and dual variables as

$$\begin{aligned} x_i(k+1) &= x_i(k) \\ &\quad - h\left(\alpha\sum_{j=1}^n L_{ij}Q(x_j(k)) + \beta v_i(k) + g_i(x_i(k))\right) \\ v_i(k+1) &= v_i(k) + h\beta\sum_{j=1}^n L_{ij}Q(x_j(k)). \end{aligned} \quad \text{(q−PDGD)}$$

Here, $x_i(0)$ are arbitrarily initialized, whereas $v_i(0) = \mathbf{0}_d$. The hyperparameters α, β and stepsize h are positive-valued. Note that only the primal quantized variable $Q(x_j(k))$ is communicated between neighbors.

Remark 1. Using Definition 3, the primal update in Algorithm (q−PDGD) can be rewritten as

$$\begin{aligned} x_i(k+1) =& x_i(k) - h\alpha\left(Q(x_i(k) - \sum_{j=1}^n a_{ij}Q(x_j(k))\right) \\ &- h\beta v_i(k) - hg_i(x_i(k)) \\ =&(1-h\alpha)x_i(k) + h\alpha\sum_{j=1}^n a_{ij}Q(x_j(k)) \\ &- h\beta v_i(k) - hg_i(x_i(k)). \end{aligned} \quad (10)$$

Upon comparing it with the DCSG algorithm from Dutta and Doan [2024]

$$\begin{aligned} x_i(k+1) =&(1-\beta_k)x_i(k) + \beta_k\sum_{j=1}^n a_{ij}Q(x_j(k)) \\ &- \alpha_k g_i(x_i(k)), \end{aligned} \quad (11)$$

we note that $h\alpha$ in (q−PDGD) is analogous to β_k in DCSG. The DCSG β-mixing term is equivalent to a Laplacian-driven consensus step in (q−PDGD), with β_k corresponding to the effective consensus step size $h\alpha$. Although both algorithms implement consensus + local innovation dynamics under quantized communication, the additional dual variable v_i in (q−PDGD) serves as a disagreement-tracking mechanism. Unlike the DCSG β_k-mixing, which explicitly controls convex averaging, v_i accumulates consensus errors and feeds them back into the primal dynamics. This feedback introduces a correction that mitigates persistent disagreement caused by mixing of quantized primal variables.

Let $\varepsilon(k)$ denote the stacked quantization noise vector such that $Q(\mathbf{x}(k)) = \mathbf{x}(k) + \varepsilon(k)$. Similarly, let $\zeta(k) = \text{col}(\zeta_1(k), \ldots, \zeta_n(k))$ denote the stacked stochastic gradients. Then, the algorithm (q−PDGD) can be written in compact stacked form as

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{x}(k) \\ &\quad - h\left(\alpha\boldsymbol{L}\mathbf{x}(k) + \beta\mathbf{v}(k) + \nabla F(\mathbf{x}(k)) + \alpha\boldsymbol{L}\varepsilon(k) + \zeta(k)\right), \end{aligned} \quad (12a)$$

$$\mathbf{v}(k+1) = \mathbf{v}(k) + h\beta\left(\boldsymbol{L}\mathbf{x}(k) + \boldsymbol{L}\varepsilon(k)\right). \quad (12b)$$

Upon defining the state $\mathbf{z}(k) = [\mathbf{x}(k)^\top, \mathbf{v}(k)^\top]^\top$, the nominal field $G(\mathbf{z})$, and total noise $\xi(k)$:

$$G(\mathbf{z}) = \begin{bmatrix} \alpha\boldsymbol{L}\mathbf{x} + \beta\mathbf{v} + \nabla F(\mathbf{x}) \\ -\beta\boldsymbol{L}\mathbf{x} \end{bmatrix}, \xi(k) = \begin{bmatrix} \alpha\boldsymbol{L}\varepsilon(k) + \zeta(k) \\ -\beta\boldsymbol{L}\varepsilon(k) \end{bmatrix},$$

the algorithm dynamics (12) becomes

$$\mathbf{z}(k+1) = \mathbf{z}(k) - h(G(\mathbf{z}(k)) + \xi(k)). \quad (13)$$

5 KEY RESULTS AND CONVERGENCE ANALYSIS OF q−PDGD

In this section, we present the convergence guarantees of the q−PDGD algorithm described in Section 4.2. We first present below the guarantees of (q−PDGD) under RSI (Assumption 3), followed by the guarantee under PL (Assumption 4).

To present the results under RSI, we introduce the following additional notations:

- $\eta = \sqrt{2}\max\left\{\frac{2\epsilon_1}{\rho_2(L)} + \alpha + 1, 4\epsilon_1 + 1\right\} > 0,$
- $\epsilon_1 = \max\left\{\frac{1}{\nu_1}\left(\frac{L_f^2}{2\beta} + \beta\rho(L)\right), \frac{1}{2}\right\},$
- $\epsilon_2 = \min\left\{\frac{\beta}{2}, \epsilon_1\nu_1\right\},$
- $\epsilon_3 = \max\left(\epsilon_1 + \frac{1}{2}, \frac{\epsilon_1}{\rho_2(L)} + \frac{\alpha}{2\beta} + \frac{1}{2}\right),$
- $\epsilon_4 = \min\left(\epsilon_1 - \frac{1}{2}, \frac{\alpha}{2\beta} - \frac{1}{2}\right),$
- $\epsilon_5 = \max\{\beta^2\rho^2(L) + 3\alpha^2\rho^2(L) + 3L_f^2, 3\beta^2\}.$
- $A = \left(\frac{\epsilon_3\eta h}{2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5}\right)\frac{nd}{4}\left(2\alpha^2 + \beta^2\right)\rho^2(L)$
- $B = \left(\frac{\epsilon_3\eta h}{2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5}\right)2n$

Throughout this section, the expectation $\mathbb{E}[\cdot]$ is taken with respect to the randomness in both the stochastic gradients and the quantization distortion of the q−PDGD algorithm, and conditioned on the history $\mathcal{F}_k$.

The following result holds for fixed stepsize h.

Theorem 1. *Let Assumptions 1-3, 5-7 hold. Consider Algorithm* (q−PDGD) *with parameters* $\alpha > \frac{2nL_f^2+\nu L_f}{\nu\rho_2(L)}$, $\beta > 0$, *and* $0 < h < \min\left\{\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}, \frac{2\epsilon_3}{\epsilon_2}\right\}$. *Then there exist* $\rho := 1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{4\epsilon_3\epsilon_4} \in (0,1)$ *and constant* $C > 0$ *such that*

$$\mathbb{E}[\|\mathbf{x}(k) - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|] \leq C\rho^k + \sqrt{A\Delta^2 + B\sigma^2}. \quad (14)$$

Theorem 1 implies that under RSI condition, for small enough but fixed step-size $h > 0$, the agents' estimates in the q−PDGD algorithm converge *linearly* in expectation to a neighborhood of some optimal solution in $x^* \in X^*$ with rate at least ρ. This neighborhood of x^* is proportional to the variance of the combined noise source. Note that exact consensus is not guaranteed above for a fixed h. However, both exact convergence to an optimal solution and consensus can be guaranteed for a diminishing step-size condition, which is formally presented in the next result.

Remark 2 (**Network scaling of** A **and** B). The constants A, B characterize the asymptotic error neighborhood in (14) arising from quantization noise (Δ) and stochastic gradient noise (σ), respectively. Appendix B elaborates how these terms are governed by the network spectral quantities $(\rho(L), \rho_2(L))$ and the network scale n while the problem parameters (L_f, ν) are $\mathcal{O}(1)$. Particularly, in the small stepsize regime, $A = \mathcal{O}\left(h\,d\,\frac{n^4\rho^4(L)}{\rho_2^3(L)}\right)$ and $B = \mathcal{O}\left(h\,\frac{n^2\rho^2(L)}{\rho_2(L)}\right)$, revealing strong sensitivity to the *network condition number* $\kappa_L := \rho(L)/\rho_2(L)$. Poor network condition (large κ_L) or large networks therefore amplify residue unless h is reduced. At the maximal stable stepsize $h = \mathcal{O}\left(\frac{\epsilon_2\,\epsilon_4}{\eta\,\epsilon_3\,\epsilon_5}\right) = \mathcal{O}\left(\frac{\rho_2^3(L)}{n^3\rho^4(L)}\right)$, the quantization-induced error is $A = \mathcal{O}(d\,n)$ and stochastic gradient-induced error $B = \mathcal{O}\left(\frac{\rho_2^2(L)}{n\,\rho^2(L)}\right)$. This behavior can be interpreted as follows. Quantization error propagates through the consensus step itself. Pushing h to its stability limit shrinks the stepsize in the exact inverse order of the dominant network amplification term, canceling spectral effects in A. However, since quantization noise is injected at every node, the dynamics accumulate n noise sources in the mixing step, yielding the unavoidable growth of A with n. On the other hand, gradient noise enters locally and are subsequently averaged by the coupled dynamics. In this case, the Laplacian acts as a variance-reduction mechanism, explaining why B increases with stronger connectivity (smaller κ_L); stronger connectivity accelerates mixing, which improves contraction but also spreads stochastic gradient noise more efficiently across the network. At the same time, larger network promotes variance averaging, yielding a decay of B with n.

Theorem 2. *Let Assumptions 1-3, 5-7 hold. Consider Algorithm* (q−PDGD) *with parameters* $\alpha > \frac{2nL_f^2+\nu L_f}{\nu\rho_2(L)}$, $\beta > 0$, *and stepsize* $h_k = \frac{\gamma}{k+1}$ *with* $\gamma > \frac{2\epsilon_3}{\epsilon_2}$. *Then there exists a constant* $C > 0$ *such that for sufficiently large* k,

$$\mathbb{E}\left[\|\mathbf{x}(k) - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|^2\right] \leq \frac{C}{k}. \quad (15)$$

Eq (15) implies that $\lim_{k\to\infty}\mathbb{E}\left[\|\mathbf{x}(k) - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|^2\right] = 0$, which means $\mathbb{E}[\|x_i(k) - \mathcal{P}_{X^*}(x_i(k))\|] \to 0$ and $\mathbb{E}[\|x_i(k) - x_j(k)\|] \to 0$ for all i, j. Thus, under RSI condition, with a diminishing stepsize, the agents' estimates in the q−PDGD algorithm achieve consensus to some optimal solution $x^* \in X^*$ in expectation at a rate of $\mathcal{O}(1/k)$, matching the best-known centralized stochastic rates in oracle complexity.

While RSI in Assumption 3 relaxes the classical strong convexity condition, an even weaker relaxation is PL in Assumption 4. We present below the convergence guarantee of Algorithm (q−PDGD) with fixed stepsize under the weaker PL condition. To present this result, we need the following

notation.

$$\epsilon \coloneqq \frac{h(\epsilon_1 - \eta\epsilon_2)}{\epsilon_5},$$
$$\kappa_1 \coloneqq \frac{1}{\rho_2(L)}\Big(4 + \frac{3}{2}L_f^2\Big),\ \kappa_2 > 1,\ \kappa_3 \coloneqq \frac{\kappa_1}{\kappa_2 - 1},$$
$$\kappa_4 \coloneqq \frac{1}{4}\Big(3 + \big(9 + 8\kappa_2 + \frac{8}{\rho_2(L)}\big)^{1/2}\Big),$$
$$\kappa_5 \coloneqq 2\Big(\kappa_2 + \frac{1}{\rho_2(L)}\Big)L_f^2 + 2\Big(\big(\kappa_2 + \frac{1}{\rho_2(L)}\big)^2 L_f^4 + L_f^2\Big)^{1/2},$$
$$\epsilon_2 \coloneqq \max\Big\{\beta^2\rho(L) + (2\alpha^2+\beta^2)\rho^2(L) + \frac{5}{2}L_f^2,\ 2\beta^2 + \frac{1}{2}\Big\},$$
$$\epsilon_3 \coloneqq \frac{1}{4} - \frac{1}{2\beta}\Big(\frac{1}{\beta} + \frac{1}{\rho_2(L)} + \frac{\alpha}{\beta}\Big)L_f^2,$$
$$\epsilon_4 \coloneqq \frac{1}{\beta^2}\Big(1 + \frac{3}{2\rho_2(L)} + \frac{3\alpha}{2\beta}\Big)L_f^2 + \frac{L_f(1+L_f)}{2},$$
$$\epsilon_5 \coloneqq \frac{\alpha+\beta}{2\beta} + \frac{1}{2\rho_2(L)},$$
$$\epsilon_1 \coloneqq \min\Big\{1, \frac{\nu}{2n}, 2\sqrt{\epsilon_2\epsilon_5}\Big\}.$$

We let $\bar{x}(k) = \frac{1}{n}(\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\boldsymbol{x}(k)$ denote the mean of agents' primal variables and $\bar{\boldsymbol{x}}(k) = \mathbf{1}_n \otimes \bar{x}(k)$ its stacked value.

Theorem 3. *Let Assumptions 1-2, 4, 6-7 hold. Consider Algorithm* (q−PDGD) *with parameters* $\beta + \kappa_1 \le \alpha \le \kappa_2\beta$, $\beta \ge \max\{\kappa_3, \kappa_4, \kappa_5\}$, *and* $0 < h < \min\Big\{\frac{\epsilon_1}{\epsilon_2}, \frac{\epsilon_3}{\epsilon_4}\Big\}$. *Then there exists a constant* $C_{err} > 0$ *such that for all agent* $i \in [n]$,

$$\limsup_{k\to\infty}\big(\mathbb{E}[\|x_{i,k} - \mathcal{P}_{X^*}(\bar{x}_k)\|^2] + \mathbb{E}[f(\bar{x}_k) - f^*]\big) \le C_{err}.$$

Consequently, under PL inequality, iterates of the q−PDGD algorithm achieve mean-square convergence in expectation to a neighborhood of the optimum. In particular, both the consensus error and objective sub-optimality gap remain uniformly bounded asymptotically.

5.1 PROOF SKETCHES OF KEY RESULTS

Proof Sketch of Theorem 1. The proof relies on constructing a composite Lyapunov function $V(\mathbf{z})$ for the augmented state $\mathbf{z} = [\mathbf{x}^\top, \mathbf{v}^\top]^\top$, defined as $V = V_2 + V_3$. Here, V_2 captures the quadratic error weighted by the spectral properties of the consensus matrix, while V_3 is a cross-term introduced to establish dissipativity and strict contraction.

First, we establish that V is bounded by squared error norm, satisfying $\epsilon_4\|\mathbf{z} - \mathbf{z}^0\|^2 \le V(\mathbf{z}) \le \epsilon_3\|\mathbf{z} - \mathbf{z}^0\|^2$. By interpreting the q−PDGD algorithm as a noisy discretization of a stable continuous-time dynamical system, we utilize the Lipschitz continuity of ∇V. We show that the deterministic drift satisfies $\langle \nabla V(\mathbf{z}), -G(\mathbf{z})\rangle \le -\frac{\epsilon_2}{\epsilon_3}V(\mathbf{z})$, which ensures exponential convergence in the noise-free case.

To account for the stochastic updates, we analyze the conditional expectation of the one-step difference $V_{k+1} - V_k$. The zero-mean property of the stochastic gradients and the quantization noise eliminates the linear noise terms. The remaining quadratic noise terms are bounded by the gradient variance σ^2 and the quantization parameter Δ^2. This yields a recursive inequality of the form

$$\mathbb{E}[V_{k+1}] \le \rho\,\mathbb{E}[V_k] + O(\Delta^2 + \sigma^2),$$

where the contraction rate $\rho \in (0, 1)$ depends on algorithm parameters. Unrolling this recursion and applying the quadratic bounds on V yields the final error bound. The detailed derivation is provided in Appendix A.3. □

Proof Sketch of Theorem 2. The proof utilizes the same V and the conditional expectation analysis from Theorem 1. By substituting the diminishing step size $h_k = \frac{\gamma}{k+1}$ into the established one-step bound, we obtain

$$\mathbb{E}[V_{k+1}] \le \left(1 - \frac{\gamma\epsilon_2}{\epsilon_3(k+1)} + O\left(\frac{1}{k^2}\right)\right)\mathbb{E}[V_k] + \frac{\bar{C}\gamma^2}{(k+1)^2},$$

where $\bar{C}$ represents the constant noise terms. For sufficiently large k, the higher-order term $O(1/k^2)$ inside the parenthesis becomes negligible compared to the linear term. Letting $\mu = \frac{\gamma\epsilon_2}{2\epsilon_3}$, the inequality simplifies to

$$\mathbb{E}[V_{k+1}] \le \left(1 - \frac{\mu}{k+1}\right)\mathbb{E}[V_k] + \frac{\bar{C}\gamma^2}{(k+1)^2}.$$

The condition $\gamma > \frac{2\epsilon_3}{\epsilon_2}$ ensures that $\mu > 1$. Solving this scalar recursion yields the decay rate $\mathbb{E}[V_k] = O(1/k)$. Combining this with the quadratic bounds on V completes the proof. See Appendix A.4 for complete details. □

Proof Sketch of Theorem 3. We employ a Lyapunov function V similar to that in Theorem 1, comprising the consensus error, the auxiliary variable tracking error, and a cross-term, but we replace the distance to the optimal solution with the objective function gap $f(\bar{x}(k)) - f^*$.

Analysis of the one-step dynamics establishes the recursion

$$\mathbb{E}[V_{k+1}] \le (1 - hc_1)\mathbb{E}[V_k] + h\tilde{C}_A,$$

where $c_1 > 0$ depends on the problem parameters and $\tilde{C}_A$ encapsulates the variance from stochastic gradients and quantization. The step size h is chosen sufficiently small such that $(1 - hc_1) \in [0, 1)$, ensuring a linear contraction of the deterministic energy. Unrolling this recursion yields

$$\mathbb{E}[V_k] \le (1 - hc_1)^k V_0 + \frac{\tilde{C}_A}{c_1}.$$

As $k \to \infty$, the term depending on the initial condition vanishes geometrically, and the expected Lyapunov value converges to a neighborhood proportional to the noise level $\tilde{C}_A$. Specifically, we obtain $\limsup_{k\to\infty}\mathbb{E}[V_k] \le \frac{\tilde{C}_A}{c_1}$, which implies the stated asymptotic bound on the consensus and objective errors. The detailed proof is in Appendix A.5. □

Remark 3. For PL, due to the coupling between the gradient tracking error and the step size, the effective noise contribution $h\tilde{C}_A$ scales as $O(h)$, unlike the $O(h^2)$ scaling observed in the RSI case. Thus, for the weaker PL condition, a diminishing stepsize does not guarantee exact convergence.

6 EXPERIMENTAL RESULTS

We corroborate the theoretical predictions from Section 5 through controlled simulations under simultaneous stochastic gradients and finite-bit random quantization. We study two nonconvex objectives. All the communication graphs in this section are considered to be a *ring*.

Example 1. We consider heterogeneous local objectives obtained by shifting a quadratic-plus-periodic function across agents, defined by

$$f_i(x) = (x - s_i)^2 + 3\sin^2(x - s_i);\ s_i := i - 1. \qquad (16)$$

This construction produces smooth but oscillatory gradients and does not rely on shared minimizers.

Example 2. We consider a piecewise-defined local objective that alternates between polynomial and curved (square-root) segments, creating multiple curvature regimes that stress stability under noise and quantization, defined by

$$f_i(x) = \begin{cases} b_{1,i}(x+1)^2, & x \le -1, \\ b_{2,i}x^4, & -1 < x \le 0, \\ 1 - \sqrt{1-x^2} + b_{3,i}x^2, & 0 < x < a, \\ \sqrt{1-(x-c)^2} - c + 1 + b_{3,i}x^2, & a \le x < 1, \\ \frac{1}{2}(x - 1 + \kappa)^2 \ + \ b_{3,i}x^2 & x \ge 1. \\ \quad + \ \sqrt{2c-2} \ + \ \dfrac{5-5c}{4}, & \end{cases} \qquad (17)$$

Here $c := \sqrt{2}$, $a := \frac{c}{2}$, $\kappa := \sqrt{\frac{c-1}{2}}$, and $b_{i,j}$ are randomly generated constants satisfying $\sum_{i=1}^n b_{i,1} > 0$, $\sum_{i=1}^n b_{i,2} = \sum_{i=1}^n b_{i,3} = 0$.

Constant vs. diminishing step-sizes. To validate the theoretical guarantees, we evaluate the proposed `q-PDGD` algorithm under both constant and diminishing step-size regimes. Figure 1 shows the consensus error and the objective gap for the non-convex problems presented in Examples 1&2 with $n = 10$ agents. With a constant step-size, the method exhibits an initial linear convergence phase and then plateaus at a noise-dominated neighborhood, consistent with Theorem 1. In contrast, applying a diminishing step-size schedule $h_k = \mathcal{O}(1/k)$ enables the agents to overcome this noise floor, achieving asymptotic consensus and driving the objective error down at a sublinear rate of $\mathcal{O}(1/k)$, corroborating Theorem 2.

Network scaling. To validate the network-dependence derived in Remark 2, we report how the total network error scales with the number of agents n while keeping other parameters fixed. The observed trend in Figure 2 matches the predicted dependence of the steady-state error on network size and spectral properties.

Baseline comparison on least squares. We benchmark the practical efficiency of `q-PDGD` against standard distributed baselines, specifically Quantized Distributed Gradient Descent `q-DGD` [Yuan et al., 2016], `CHOCO-SGD` [Koloskova* et al., 2020] and `DCSG` [Dutta and Doan, 2024]. To establish a fair baseline against `q-DGD` and `DCSG`, which typically require stronger assumptions like strong convexity, we evaluate the algorithms on a standard distributed least squares problem. This fair comparison isolates and highlights the accelerated convergence of `q-PDGD`. In this problem, each agent's local cost is defined by

$$f_i(x_i) = \frac{1}{2}\|A_i x_i - b_i\|^2.$$

We simulate $n = 10$ agents estimating a signal $x^* \in \mathbb{R}^3$ ($d = 3$). The data matrices $A_i \in \mathbb{R}^{3\times 3}$ and x^* are drawn i.i.d. from standard normal distribution $\mathcal{N}(0,1)$, with measurements $b_i = A_i x^*$ ensuring a true minimum $f^* = 0$. Stochastic local gradients are computed as $g_i(x) = A_i^T(A_i x_i - b_i) + \zeta_i$, where $\zeta_i \sim \mathcal{N}(0, 0.5)$. Communication is compressed using an 8-bit random uniform quantizer over $[-10, 10]$. For this least-squares problem, Figure 3 showcases the objective gap trajectories under simultaneous stochastic gradient noise and random quantization. `q-PDGD` achieves the target optimality threshold of 0.01 at $k = 215$, significantly outperforming `q-DGD` ($k = 745$), `CHOCO-SGD` ($k = 948$) and `DCSG` ($k = 969$). As further highlighted, `q-PDGD` exhibits remarkable robustness to finite-bit communication; its quantized trajectory closely mirrors its unquantized (exact communication) counterpart. We further stress-test `q-PDGD` against representative quantized decentralized baselines on two image-classification benchmarks where our assumptions are formally violated. As shown in Appendix D, the consensus advantage predicted by Theorem 1 continues to hold empirically. This confirms that the primal-dual tracking mechanism effectively absorbs and mitigates the consensus errors induced by random quantization without severely degrading the convergence rate. Additional plots demonstrating the dependency of final error on Δ^2 and σ^2 from Theorem 1 and performance comparison of `q-PDGD` with the aforementioned baselines to reach a target threshold are included in Appendix C.

7 CONCLUSION

In summary, we developed `q-PDGD`, a stochastic primal–dual method that remains provably effective under simultaneous gradient noise and finite-bit, random (unbiased) quantized communication in fully decentralized networks. Under RSI, we established a clear linear-to-neighborhood

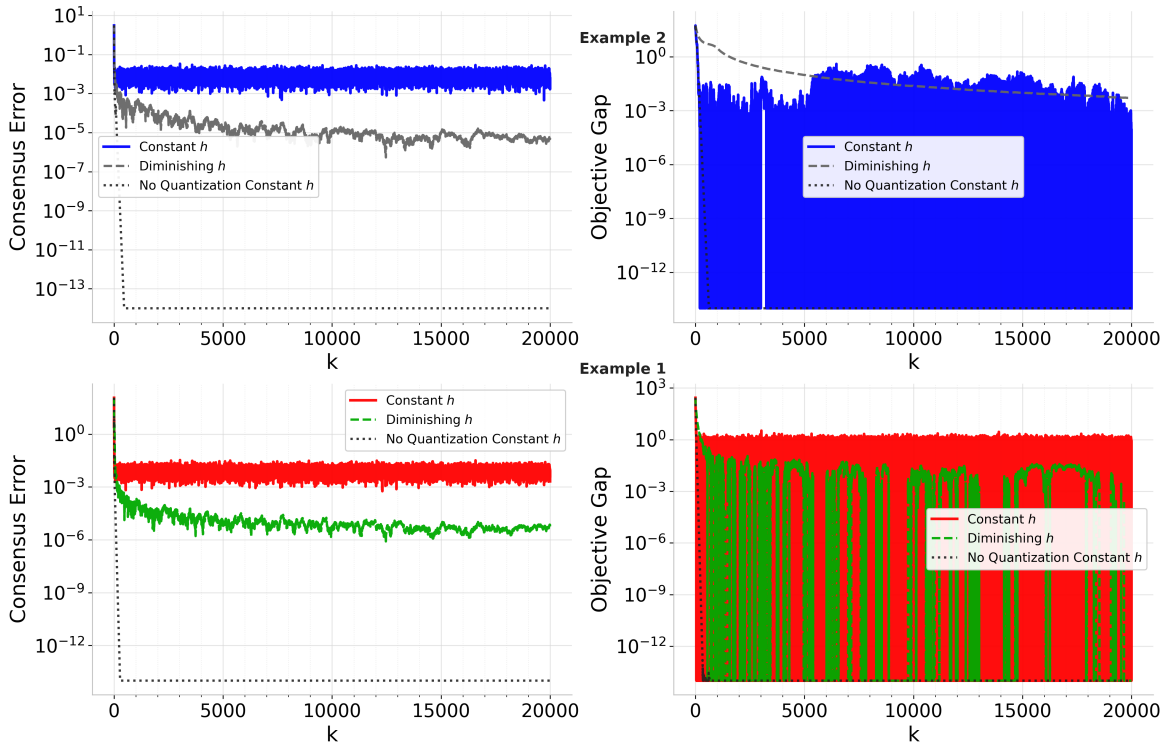

Figure 1: **(Example 2, top) and (Example 1, bottom).** Evolution of the consensus error and objective gap for q−PDGD under constant and diminishing step-sizes over a ring graph with $n = 10$ agents. Constant step-sizes converge linearly before plateauing at a neighborhood, while diminishing step-size achieves consensus and optimization.

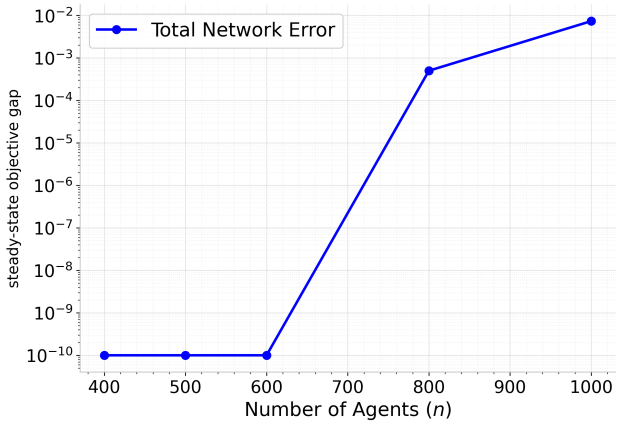

Figure 2: Total network error scaling versus the number of agents n for Example 2, validating the dependency predicted by Remark 2.

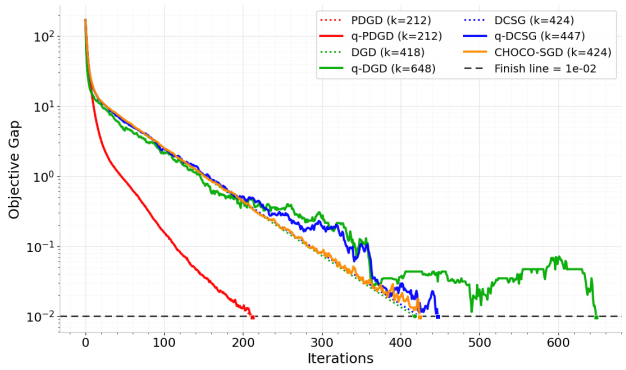

Figure 3: Quantized (Q) versus unquantized (UQ) trajectories on distributed least squares under stochastic gradients and finite-bit random quantization. q−PDGD remains robust to compression.

guarantee for constant step-sizes with an explicit residual radius that cleanly decomposes the contributions of quantization distortion, stochastic variance, and graph connectivity, and we further showed that a diminishing step-size achieves the minimax-optimal $\mathcal{O}(1/k)$ rate without shared-minimizer assumptions. Under the even weaker PL geometry, we proved linear convergence to a noise-dominated neighborhood, again explicitly characterizing how network structure and compression affect the steady-state error. Experiments corroborate these predictions by exhibiting the expected plateau vs. $\mathcal{O}(1/k)$ behavior and by validating the linear dependence on (Δ^2) and (σ^2) as well as the predicted scaling with network size and spectral properties. Collectively, these results position q−PDGD as a principled and practical tool for communication-constrained distributed learning, and they motivate future directions such as sharper constants, adaptive bit/step-size scheduling, and extensions to biased compressors and time-varying graphs.

References

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30, 2017.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed Optimisation for Non-Convex Problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

Ziqin Chen, Shu Liang, Li Li, and Shuming Cheng. Quantized Primal-Dual Algorithms for Network Optimization with Linear Convergence. *IEEE Transactions on Automatic Control*, 69(1):471–478, 2023.

Chee-Yee Chong, Kuo-Chu Chang, and Shozo Mori. A Review of Forty Years of Distributed Estimation. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2018.

Amit Dutta and Thinh T Doan. On the O(1/k) Convergence of Distributed Gradient Methods Under Random Quantization. *IEEE Control Systems Letters*, 8:2967–2972, 2024.

Majid Jahani, Xi He, Chenxin Ma, Aryan Mokhtari, Dheevatsa Mudigere, Alejandro Ribeiro, and Martin Takác. Efficient Distributed Hessian Free Algorithm for Large-Scale Empirical Risk Minimization via Accumulating Sample Strategy. In *International Conference on Artificial Intelligence and Statistics*, pages 2634–2644. PMLR, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak-Łojasiewicz Condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Solmaz S Kia, Jorge Cortés, and Sonia Martínez. Distributed convex Optimization via Continuous-Time Coordination Algorithms with Discrete-Time Communication. *Automatica*, 55:254–264, 2015.

Anastasia Koloskova*, Tao Lin*, Sebastian U Stich, and Martin Jaggi. Decentralized Deep Learning with Arbitrary Communication Compression. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkgGCkrKvH`.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed Learning with Compressed Gradient Differences. *Optimization Methods and Software*, 40(5):1181–1196, 2025.

Angelia Nedić and Ji Liu. Distributed Optimization for Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):77–103, 2018.

Ye Pu, Melanie N Zeilinger, and Colin N Jones. Quantization Design for Distributed Optimization. *IEEE Transactions on Automatic Control*, 62(5):2107–2120, 2016.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Quantized Decentralized Consensus Optimization. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5838–5843. IEEE, 2018.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Apostolos I Rikos, Wei Jiang, Themistoklis Charalambous, and Karl H Johansson. Distributed Optimization with Gradient Descent and Quantized Communication. *IFAC-PapersOnLine*, 56(2):5900–5906, 2023.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An Exact First-Order Algorithm for Decentralized Consensus Optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A Survey of Distributed Optimization. *Annual Reviews in Control*, 47:278–305, 2019.

Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Exponential Convergence for Distributed Optimization under the Restricted Secant Inequality Condition. *IFAC-PapersOnLine*, 53(2):2672–2677, 2020a.

Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Linear Convergence for Distributed Optimization without Strong Convexity. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3643–3648. IEEE, 2020b.

Wenrui Yu, Qiongxiu Li, Milan Lopuhaä-Zwakenberg, Mads Græsbøll Christensen, and Richard Heusdens. Provable Privacy Advantages of Decentralized Federated Learning via Distributed Optimization. *IEEE Transactions on Information Forensics and Security*, 20:822–838, 2024.

Yue Yu, Jiaxiang Wu, and Longbo Huang. Double Quantization for Communication-Efficient Distributed Optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Kun Yuan, Qing Ling, and Wotao Yin. On the Convergence of Decentralized Gradient Descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016. doi: 10.1137/130943170. URL https://doi.org/10.1137/130943170.

Shengyu Zhu, Mingyi Hong, and Biao Chen. Quantized Consensus ADMM for Multi-Agent Distributed Optimization. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4134–4138. IEEE, 2016.

Quantized Stochastic Primal–Dual Methods for Distributed Optimization under Relaxed Global Geometry (Supplementary Material)

Susmit Sarkar[1,*] **Abhinav Raghuvanshi**[2,*] **Kushal Chakrabarti**[2] **Mayank Baranwal**[1,2]

[1]Indian Institute of Technology Bombay, Mumbai, Maharashtra, India
[2]Tata Consultancy Services Research, Thane, Maharashtra, India
[*]Equal Contribution

A DETAILED PROOFS

A.1 MATRIX NOTATION AND SPECTRAL PROPERTIES

Recall L denote the Laplacian matrix of the connected undirected graph $\mathcal{G}$. Let the *centering matrix* K_n denote the orthogonal projection onto the subspace orthogonal to $\mathbf{1}_n$, defined as

$$K_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top. \tag{18}$$

Both L and K_n are positive semi-definite. Since the graph is connected under Assumption 6, the null spaces satisfy $\text{null}(L) = \text{null}(K_n) = \text{span}(\mathbf{1}_n)$. Furthermore, the matrices satisfy the commutativity and projection properties

$$K_nL = LK_n = L, \quad \text{and} \quad \rho(K_n) = 1. \tag{19}$$

Let $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of L. We denote the spectral radius by $\rho(L) = \lambda_n$ and the algebraic connectivity (smallest non-zero eigenvalue) by $\rho_2(L) = \lambda_2$. The Laplacian satisfies the following bounds relative to K_n:

$$\rho_2(L)K_n \preceq L \preceq \rho(L)K_n. \tag{20}$$

To analyze the weighted norm in the Lyapunov function, we introduce the spectral decomposition of L. There exists an orthogonal matrix $Q = [r, R] \in R^{n\times n}$, where $r = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $R \in R^{n\times(n-1)}$, such that [Yi et al., 2020a]

$$L = [r\, R]\begin{bmatrix}0 & 0\\ 0 & \Lambda_1\end{bmatrix}\begin{bmatrix}r^\top\\ R^\top\end{bmatrix} = R\Lambda_1R^\top, \tag{21}$$

$$R\Lambda_1^{-1}R^\top L = LR\Lambda_1^{-1}R^\top = K_n, \tag{22}$$

$$\frac{1}{\rho(L)}K_n \preceq R\Lambda_1^{-1}R^\top \preceq \frac{1}{\rho_2(L)}K_n. \tag{23}$$

where $\Lambda_1 = \text{diag}([\lambda_2, \ldots, \lambda_n])$.

A.2 PROOF OF LEMMA 2

Let $\mathbf{x}^* = \mathcal{P}_{\mathcal{X}^*}(\mathbf{x})$. For any $\mathbf{x} \in \mathbb{R}^{nd}$, we orthogonally decompose it as $\mathbf{x} = \mathbf{a} + \mathbf{b}$, where $\mathbf{a} \in \mathbb{H} = \{\mathbf{1}_n \otimes y : y \in \mathbb{R}^d\}$ is the average component, and $\mathbf{b} \in \mathbb{H}^\perp$ is the disagreement component. Specifically, $\mathbf{a} = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\mathbf{x}$.

Since $\mathcal{X}^* \subseteq \mathbb{H}$, the projection onto the optimal set is determined solely by the average component, i.e., $\mathcal{P}_{\mathcal{X}^*}(\mathbf{x}) = \mathcal{P}_{\mathcal{X}^*}(\mathbf{a}) = \mathbf{x}^*$. Consequently, the error norm decomposes as

$$\|\mathbf{x} - \mathbf{x}^*\|^2 = \|\mathbf{a} - \mathbf{x}^*\|^2 + \|\mathbf{b}\|^2. \tag{24}$$

We expand the inner product term involving the gradients as

$$\begin{aligned}&(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{x} - \mathbf{x}^*) \\ &= (\nabla F(\mathbf{a}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{a} - \mathbf{x}^*) + (\nabla F(\mathbf{x}) - \nabla F(\mathbf{a}))^\top (\mathbf{a} - \mathbf{x}^*) + (\nabla F(\mathbf{a}) - \nabla F(\mathbf{x}^*))^\top \mathbf{b} \\ &+ (\nabla F(\mathbf{x}) - \nabla F(\mathbf{a}))^\top \mathbf{b}. \end{aligned} \tag{25}$$

We bound these terms using the RSI Condition. Since $\mathbf{a}, \mathbf{x}^* \in \mathbb{H}$, we can relate F to the global cost f. From Assumption 3,

$$(\nabla F(\mathbf{a}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{a} - \mathbf{x}^*) = (\nabla f(a) - \nabla f(x^*))^\top (a - x^*) \geq \frac{\nu}{n} \|\mathbf{a} - \mathbf{x}^*\|^2. \tag{26}$$

From Lipschitz continuity in Assumption 2,

$$(\nabla F(\mathbf{x}) - \nabla F(\mathbf{a}))^\top (\mathbf{a} - \mathbf{x}^*) \geq -L_f \|\mathbf{x} - \mathbf{a}\| \|\mathbf{a} - \mathbf{x}^*\| = -L_f \|\mathbf{b}\| \|\mathbf{a} - \mathbf{x}^*\|, \tag{27}$$

$$(\nabla F(\mathbf{x}) - \nabla F(\mathbf{a}))^\top \mathbf{b} \geq -L_f \|\mathbf{x} - \mathbf{a}\| \|\mathbf{b}\| = -L_f \|\mathbf{b}\|^2, \tag{28}$$

$$(\nabla F(\mathbf{a}) - \nabla F(\mathbf{x}^*))^\top \mathbf{b} \geq -L_f \|\mathbf{a} - \mathbf{x}^*\| \|\mathbf{b}\|. \tag{29}$$

Combining the inequalities above,

$$(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{x} - \mathbf{x}^*) \geq \frac{\nu}{n} \|\mathbf{a} - \mathbf{x}^*\|^2 - 2L_f \|\mathbf{b}\| \|\mathbf{a} - \mathbf{x}^*\| - L_f \|\mathbf{b}\|^2. \tag{30}$$

Using Young's inequality, we get

$$-2L_f \|\mathbf{b}\| \|\mathbf{a} - \mathbf{x}^*\| \geq -\frac{\nu}{2n} \|\mathbf{a} - \mathbf{x}^*\|^2 - \frac{2nL_f^2}{\nu} \|\mathbf{b}\|^2.$$

Substituting above in (30),

$$(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{x} - \mathbf{x}^*) \geq \frac{\nu}{2n} \|\mathbf{a} - \mathbf{x}^*\|^2 - \left(\frac{2nL_f^2}{\nu} + L_f \right) \|\mathbf{b}\|^2. \tag{31}$$

Now consider the Laplacian term. Since $\mathbf{x}^\top \boldsymbol{L} \mathbf{x} = \mathbf{b}^\top \boldsymbol{L} \mathbf{b}$ and $L \succeq \rho_2(L) K_n$ on the subspace $\mathbb{H}^\perp$,

$$\alpha \mathbf{x}^\top \boldsymbol{L} \mathbf{x} \geq \alpha \, \rho_2(L) \|\mathbf{b}\|^2. \tag{32}$$

Adding (31) and (32),

$$\begin{aligned}&(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*))^\top (\mathbf{x} - \mathbf{x}^*) + \alpha \mathbf{x}^\top \boldsymbol{L} \mathbf{x} \\ &\geq \frac{\nu}{2n} \|\mathbf{a} - \mathbf{x}^*\|^2 + \left(\alpha \rho_2(L) - \frac{2nL_f^2}{\nu} - L_f \right) \|\mathbf{b}\|^2 \\ &\geq \min \left\{ \frac{\nu}{2n}, \alpha \rho_2(L) - \frac{2nL_f^2 + \nu L_f}{\nu} \right\} (\|\mathbf{a} - \mathbf{x}^*\|^2 + \|\mathbf{b}\|^2) \\ &= \nu_1 \|\mathbf{x} - \mathbf{x}^*\|^2. \end{aligned} \tag{33}$$

Given the condition on α, the coefficient above is strictly positive, ensuring $\nu_1 > 0$.

A.3 PROOF OF THEOREM 1

Consider the following functions

$$V_1(\mathbf{x}, \mathbf{v}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}^0\|^2 + \frac{1}{2} \|\mathbf{v} - \mathbf{v}^0\|^2_{R\Lambda_1^{-1} R^\top \otimes \boldsymbol{I}_d}, \tag{34}$$

$$V_2(\mathbf{x}, \mathbf{v}) = 2\epsilon_1 V_1(\mathbf{x}, \mathbf{v}) + \frac{\alpha}{2\beta} \|\mathbf{v} - \mathbf{v}^0\|^2, \tag{35}$$

$$V_3(\mathbf{x}, \mathbf{v}) = \mathbf{x}^\top (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0), \tag{36}$$

where, R and Λ are defined in Section A.1, $\mathbf{x}^0 = \mathbf{1}_n \otimes x^0 = \mathcal{P}_{\mathcal{X}^*}(\mathbf{x})$, and $\mathbf{v}^0 = -\frac{1}{\beta}\nabla f(\mathbf{x}^0)$. Consider the Lyapunov function candidate

$$V(\mathbf{x}, \mathbf{v}) = V_2(\mathbf{x}, \mathbf{v}) + V_3(\mathbf{x}, \mathbf{v}). \tag{37}$$

Since $\mathbf{1}_n^\top L = \mathbf{0}_n^\top$, the v_i-update in (q-PDGD) implies that the sum of the auxiliary variables is invariant with respect to k, i.e., $\sum_{i=1}^n v_i(k+1) = \sum_{i=1}^n v_i(k)$. So, given the initialization $\sum_{i=1}^n v_i(0) = \mathbf{0}_d$, we have $\sum_{i=1}^n v_i(k) = \mathbf{0}_p$ for all $k \geq 0$. This implies that the vector $\mathbf{v}$ lies entirely in the range space of the centering matrix $(K_n \otimes \boldsymbol{I}_d)$. Furthermore, since $\mathbf{v}^0 = -\frac{1}{\beta}\nabla f(\mathbf{x}^0)$ and $\sum_{i=1}^n \nabla f_i(x^0) = \mathbf{0}_p$, we have $(K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0) = \mathbf{v} - \mathbf{v}^0$.

Recall from (23) that the matrix $R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d$ satisfies the following bounds:

$$\frac{1}{\rho(L)}(K_n \otimes \boldsymbol{I}_d) \preceq R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d \preceq \frac{1}{\rho_2(L)}(K_n \otimes \boldsymbol{I}_d). \tag{38}$$

Using above, we bound the second term in V_1 as

$$\|\mathbf{v} - \mathbf{v}^0\|^2_{R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d} \leq \frac{1}{\rho_2(L)}(\mathbf{v} - \mathbf{v}^0)^\top (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0) = \frac{1}{\rho_2(L)}\|\mathbf{v} - \mathbf{v}^0\|^2.$$

Next, we bound the cross-term V_3 using Young's inequality. Noting that $\mathbf{x}^0 = \mathbf{1}_n \otimes x^0$, we have $(\mathbf{x}^0)^\top (K_n \otimes \boldsymbol{I}_d) = \mathbf{0}_{nd}$. Thus, V_3 can be rewritten as

$$V_3(\mathbf{x}, \mathbf{v}) = (\mathbf{x} - \mathbf{x}^0)^\top (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0).$$

Using Young's inequality and $\|K_n \otimes \boldsymbol{I}_d\| = 1$, we obtain

$$-\frac{1}{2}(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2) \leq V_3(\mathbf{x}, \mathbf{v}) \leq \frac{1}{2}(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2).$$

Substituting these bounds into the definition of $V(\mathbf{x}, \mathbf{v})$ in (37),

$$\begin{aligned} V &\leq 2\epsilon_1 \left(\frac{1}{2}\|\mathbf{x} - \mathbf{x}^0\|^2 + \frac{1}{2\rho_2(L)}\|\mathbf{v} - \mathbf{v}^0\|^2\right) + \frac{\alpha}{2\beta}\|\mathbf{v} - \mathbf{v}^0\|^2 + \frac{1}{2}(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2) \\ &\leq \epsilon_3(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2), \end{aligned} \tag{39}$$

where $\epsilon_3 = \max\left(\epsilon_1 + \frac{1}{2}, \frac{\epsilon_1}{\rho_2(L)} + \frac{\alpha}{2\beta} + \frac{1}{2}\right)$. Similarly, for the lower bound of V, we can obtain

$$\begin{aligned} V &\geq 2\epsilon_1 \left(\frac{1}{2}\|\mathbf{x} - \mathbf{x}^0\|^2\right) + \frac{\alpha}{2\beta}\|\mathbf{v} - \mathbf{v}^0\|^2 - \frac{1}{2}(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2) \\ &\geq \epsilon_4(\|\mathbf{x} - \mathbf{x}^0\|^2 + \|\mathbf{v} - \mathbf{v}^0\|^2), \end{aligned} \tag{40}$$

where $\epsilon_4 = \min\left(\epsilon_1 - \frac{1}{2}, \frac{\alpha}{2\beta} - \frac{1}{2}\right)$. By choosing, $\epsilon_1 > \frac{1}{2}$ and $\frac{\alpha}{\beta} > 1$, we ensure $\epsilon_4 > 0$. Thus,

$$\epsilon_4(\|\mathbf{v} - \mathbf{v}^0\|^2 + \|\mathbf{x} - \mathbf{x}^0\|^2) \leq V(\mathbf{z}) \leq \epsilon_3(\|\mathbf{v} - \mathbf{v}^0\|^2 + \|\mathbf{x} - \mathbf{x}^0\|^2). \tag{41}$$

Recall the state vector $\mathbf{z}(k) = [\mathbf{x}(k)^\top, \mathbf{v}(k)^\top]^\top$. The Lyapunov function $V(\mathbf{z})$ is differentiable and its gradient is $\nabla V(\mathbf{z}) = [\nabla V(\mathbf{z})_1^\top, \nabla V(\mathbf{z})_2^\top]^\top$, where

$$\begin{aligned} [\nabla V(\mathbf{z})]_1 &= 2\epsilon_1(\mathbf{x} - \mathbf{x}^0) + (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0), \\ [\nabla V(\mathbf{z})]_2 &= (K_n \otimes \boldsymbol{I}_d)\mathbf{x} + \left(2\epsilon_1(R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d) + \frac{\alpha}{\beta} \otimes \boldsymbol{I}_{nd}\right)(\mathbf{v} - \mathbf{v}^0). \end{aligned}$$

Also, $\nabla V(\mathbf{z})$ is Lipschitz continuous with constant $\eta = \sqrt{2}\max\left\{\frac{2\epsilon_1}{\rho_2(L)} + \alpha + 1, 4\epsilon_1 + 1\right\} > 0$. Under Assumption 2 using the descent lemma for smooth functions, and substituting from (13),

$$\begin{aligned} &V(\mathbf{z}(k+1)) - V(\mathbf{z}(k)) \leq \langle \nabla V(\mathbf{z}(k)), \mathbf{z}(k+1) - \mathbf{z}(k)\rangle + \frac{\eta}{2}\|\mathbf{z}(k+1) - \mathbf{z}(k)\|^2 \\ &= -h\langle \nabla V(\mathbf{z}(k)), G(\mathbf{z}(k))\rangle - h\langle \nabla V(\mathbf{z}(k)), \xi(k) + \frac{\eta h^2}{2}(\|G(\mathbf{z}(k))\|^2 + \|\xi(k)\|^2 \\ &+ 2\langle G(\mathbf{z}(k)), \xi(k)\rangle). \end{aligned} \tag{42}$$

To derive a bound for the deterministic term $-\langle \nabla V, F\rangle$, we observe that $-\langle \nabla V(\mathbf{z}), G(\mathbf{z})\rangle$ corresponds to the time-derivative $\dot{V}$ of the same Lyapunov function along the trajectories of only the deterministic component of the continuous-time counterpart of (13), defined by

$$\dot{\mathbf{x}}(t) = -\alpha \boldsymbol{L}\mathbf{x}(t) - \beta \mathbf{v}(t) - \nabla F(\mathbf{x}(t)), \tag{43a}$$
$$\dot{\mathbf{v}}(t) = \beta \boldsymbol{L}\mathbf{x}(t). \tag{43b}$$

We have, $\mathbf{1}^\top \mathbf{v}(t) = \mathbf{1}^\top \mathbf{v}(0) = \mathbf{0}_{nd}$. Considering the projection matrix $P = K_n \otimes \boldsymbol{I}_d$, we have

$$P(\mathbf{v} - \mathbf{v}^0) = \mathbf{v} - \mathbf{v}^0.$$

Recalling $V_1(\mathbf{x}, \mathbf{v}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^0\|^2 + \frac{1}{2}\|\mathbf{v} - \mathbf{v}^0\|^2_{R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d}$, the derivative is

$$\begin{aligned}\dot{V}_1 &= (\mathbf{x} - \mathbf{x}^0)^\top \dot{\mathbf{x}} + (\mathbf{v} - \mathbf{v}^0)^\top (R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d)\dot{\mathbf{v}} \\ &= (\mathbf{x} - \mathbf{x}^0)^\top (-\alpha \boldsymbol{L}\mathbf{x} - \beta \mathbf{v} - \nabla F(\mathbf{x})) + (\mathbf{v} - \mathbf{v}^0)^\top (R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d)(\beta \boldsymbol{L}\mathbf{x}).\end{aligned} \tag{44}$$

Using $(R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_d)\boldsymbol{L} = K_n \otimes \boldsymbol{I}_d = P$, $\boldsymbol{L}\mathbf{x}^0 = \mathbf{0}$, $P(\mathbf{v} - \mathbf{v}^0) = \mathbf{v} - \mathbf{v}^0$, and Lemma 2,

$$\dot{V}_1 = -\alpha \mathbf{x}^\top \boldsymbol{L}\mathbf{x} - (\mathbf{x} - \mathbf{x}^0)^\top (\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^0)) \leq -\nu_1 \|\mathbf{x} - \mathbf{x}^0\|^2. \tag{45}$$

Differentiating $V_2(\mathbf{z}) = 2\epsilon_1 V_1(\mathbf{z}) + \frac{\alpha}{2\beta}\|\mathbf{v} - \mathbf{v}^0\|^2$,

$$\begin{aligned}\dot{V}_2 = 2\epsilon_1 \dot{V}_1 + \frac{\alpha}{\beta}(\mathbf{v} - \mathbf{v}^0)^\top \dot{\mathbf{v}} &= 2\epsilon_1 \dot{V}_1 + \frac{\alpha}{\beta}(\mathbf{v} - \mathbf{v}^0)^\top (\beta \boldsymbol{L}\mathbf{x}) \\ &\leq -2\epsilon_1 \nu_1 \|\mathbf{x} - \mathbf{x}^0\|^2 + \alpha(\mathbf{v} - \mathbf{v}^0)^\top \boldsymbol{L}\mathbf{x}.\end{aligned} \tag{46}$$

Differentiating $V_3(\mathbf{z}) = (\mathbf{x} - \mathbf{x}^0)^\top (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0)$,

$$\begin{aligned}\dot{V}_3 &= (\mathbf{x} - \mathbf{x}^0)^\top (K_n \otimes \boldsymbol{I}_d)\dot{\mathbf{v}} + \dot{\mathbf{x}}^\top (K_n \otimes \boldsymbol{I}_d)(\mathbf{v} - \mathbf{v}^0) \\ &\leq \beta\rho(L)\|\mathbf{x} - \mathbf{x}^0\|^2 - \alpha(\mathbf{v} - \mathbf{v}^0)^\top \boldsymbol{L}\mathbf{x} - \beta\|\mathbf{v} - \mathbf{v}^0\|^2 + \frac{L_f^2}{2\beta}\|\mathbf{x} - \mathbf{x}^0\|^2 + \frac{\beta}{2}\|\mathbf{v} - \mathbf{v}^0\|^2 \\ &= -\alpha(\mathbf{v} - \mathbf{v}^0)^\top \boldsymbol{L}\mathbf{x} - \frac{\beta}{2}\|\mathbf{v} - \mathbf{v}^0\|^2 + \left(\beta\rho(L) + \frac{L_f^2}{2\beta}\right)\|\mathbf{x} - \mathbf{x}^0\|^2.\end{aligned} \tag{47}$$

Adding (46) and (47),

$$\dot{V} = \dot{V}_2 + \dot{V}_3 \leq \left(-2\epsilon_1\nu_1 + \beta\rho(L) + \frac{L_f^2}{2\beta}\right)\|\mathbf{x} - \mathbf{x}^0\|^2 - \frac{\beta}{2}\|\mathbf{v} - \mathbf{v}^0\|^2. \tag{48}$$

By choosing $\epsilon_1 \geq \frac{1}{\nu_1}\left(\beta\rho(L) + \frac{L_f^2}{2\beta}\right)$, we have

$$\dot{V} \leq -\epsilon_1\nu_1\|\mathbf{x} - \mathbf{x}^0\|^2 - \frac{\beta}{2}\|\mathbf{v} - \mathbf{v}^0\|^2. \tag{49}$$

Define $\epsilon_2 = \min\left\{\epsilon_1\nu_1, \frac{\beta}{2}\right\}$. From the upper bound in (41) and above, we conclude $\dot{V} \leq -\frac{\epsilon_2}{\epsilon_3}V$. We have obtained that the first term in the RHS of (42) is bounded by

$$-F^\top(\mathbf{z}(k))\nabla V(\mathbf{z}(k)) = \dot{V}(\mathbf{z}(k)) \leq -\frac{\epsilon_2}{\epsilon_3}V(\mathbf{z}(k)). \tag{50}$$

Next, we upper bound the other deterministic term $\|F\|^2$ in (42). Using $\boldsymbol{L}\mathbf{x}^0 = \mathbf{0}_{nd}$ and $\mathbf{v}^0 = -\frac{1}{\beta}\nabla F(\mathbf{x}^0)$, we rewrite

$$G(\mathbf{z}) = \begin{bmatrix}\alpha\boldsymbol{L}(\mathbf{x} - \mathbf{x}^0) + \beta(\mathbf{v} - \mathbf{v}^0) + \nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^0) \\ -\beta\boldsymbol{L}(\mathbf{x} - \mathbf{x}^0)\end{bmatrix}.$$

Using the triangle inequality, Lipschitz continuity of ∇F from Assumption 2, and Section A.1,

$$\begin{aligned}\|G(\mathbf{z})\|^2 &\leq 3\|\alpha\boldsymbol{L}(\mathbf{x}-\mathbf{x}^0)\|^2 + 3\|\beta(\mathbf{v}-\mathbf{v}^0)\|^2 + 3\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^0)\|^2 + \|\beta\boldsymbol{L}(\mathbf{x}-\mathbf{x}^0)\|^2 \\ &\leq (\beta^2\rho^2(L) + 3\alpha^2\rho^2(L) + 3L_f^2)\|\mathbf{x}-\mathbf{x}^0\|^2 + 3\beta^2\|\mathbf{v}-\mathbf{v}^0\|^2 \\ &\leq \epsilon_5(\|\mathbf{x}-\mathbf{x}^0\|^2 + \|\mathbf{v}-\mathbf{v}^0\|^2),\end{aligned}$$

where $\epsilon_5 = \max\{3\alpha^2\rho^2(L) + \beta^2\rho^2(L) + 3L_f^2, 3\beta^2\}$. Using the lower bound from (41), we obtain

$$\|G(\mathbf{z}(k))\|^2 \leq \frac{\epsilon_5}{\epsilon_4}V(\mathbf{z}(k)). \tag{51}$$

Upon substituting from above in (42),

$$\begin{aligned}&V(\mathbf{z}(k+1)) \\ &\leq V(\mathbf{z}(k)) - \frac{h\epsilon_2}{\epsilon_3}V(\mathbf{z}(k)) + \frac{\eta h^2\epsilon_5}{2\epsilon_1}V(\mathbf{z}(k)) - h\langle\nabla V(\mathbf{z}(k)),\xi(k)\rangle + \frac{\eta h^2}{2}(\|\xi(k)\|^2 + 2\langle G(\mathbf{z}(k)),\xi(k)\rangle) \\ &= (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})V(\mathbf{z}(k)) - h\langle\nabla V(\mathbf{z}(k)),\xi(k)\rangle + \frac{\eta h^2}{2}(\|\xi(k)\|^2 + 2\langle G(\mathbf{z}(k)),\xi(k)\rangle).\end{aligned}$$

Taking conditional expectation with respect to $\mathcal{F}_k$, and using the fact that the random quantization model and stochastic gradients are unbiased (Assumption 7 and (5)), we have

$$\begin{aligned}\mathbb{E}[V(\mathbf{z}(k+1)) \mid \mathcal{F}_k] &\leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})V(z(k)) - h\mathbb{E}[\langle\nabla V(\mathbf{z}(k)),\xi(k)\rangle \mid \mathcal{F}_k] \\ &\quad + \eta h^2\mathbb{E}[\langle G(\mathbf{z}(k)),\xi(k)\rangle \mid \mathcal{F}_k] + \frac{\eta h^2}{2}\mathbb{E}[\|\xi(k)\|^2 \mid \mathcal{F}_k] \\ \implies \mathbb{E}[V(\mathbf{z}(k+1))] &\leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})\mathbb{E}[V(\mathbf{z}(k))] + \frac{\eta h^2}{2}\mathbb{E}[\|\xi(k)\|^2],\end{aligned}$$

as $\mathbb{E}[\xi(k) \mid \mathcal{F}_k] = \mathbf{0}_{2nd}$. Moreover, from Assumption 7 and (5), it follows that

$$\mathbb{E}[\|\xi(k)\|^2] = \mathbb{E}[\|\alpha\mathcal{L}\varepsilon(k) + \zeta(k)\|^2] + \mathbb{E}[\|\beta\mathcal{L}\varepsilon(k)\|^2] \leq (2\alpha^2 + \beta^2)\rho^2(L)\mathbb{E}[\|\varepsilon(k)\|^2] + 2\mathbb{E}[\|\zeta(k)\|^2],$$

so that

$$\mathbb{E}[\|\xi(k)\|^2] \leq (2\alpha^2 + \beta^2)\rho^2(L)\frac{nd\Delta^2}{4} + 2n\sigma^2$$

Combining these bounds yields

$$\begin{aligned}\mathbb{E}[V(\mathbf{z}(k+1))] &\leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})\mathbb{E}[V(\mathbf{z}(k))] + \frac{\eta h^2}{2}\left((2\alpha^2 + \beta^2)\rho^2(L)\frac{nd\Delta^2}{4} + 2n\sigma^2\right) \\ &= (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})\mathbb{E}[V(\mathbf{z}(k))] + \frac{\eta h^2}{2}(C_1\Delta^2 + C_2\sigma^2).\end{aligned} \tag{52}$$

Since h satisfies $0 < h < \min\left\{\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}, \frac{2\epsilon_3}{\epsilon_2}\right\}$, we have $0 < (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4}) < 1$. Then, unrolling the above recursion from k to 0,

$$\begin{aligned}\mathbb{E}[V(\mathbf{z}(k))] &\leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})^k\mathbb{E}[V(\mathbf{z}(0))] + \frac{\epsilon_3\epsilon_4\eta h}{(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}(C_1\Delta^2 + C_2\sigma^2) \\ &= (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})^k\mathbb{E}[V(\mathbf{z}(0))] + K_1\Delta^2 + K_2\sigma^2.\end{aligned}$$

Then, from (41) the definition of $\mathbf{x}^0$,

$$\mathbb{E}[\|\mathbf{x}(k) - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|^2] \leq \frac{\mathbb{E}[V(\mathbf{z}(k))]}{\epsilon_4} \leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})^k\frac{\mathbb{E}[V(\mathbf{z}(0))]}{\epsilon_4} + A\Delta^2 + B\sigma^2.$$

Thus,

$$\mathbb{E}[\|\mathbf{x}(k) - \mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|] \leq (1 - \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{4\epsilon_3\epsilon_4})^k\sqrt{\frac{\mathbb{E}[V(\mathbf{z}(0))]}{\epsilon_4}} + \sqrt{A\Delta^2 + B\sigma^2}.$$

The proof is complete.

A.4 PROOF OF THEOREM 2

Consider the same Lyapunov function

$$V(\mathbf{z}) = V_2(\mathbf{x}, \mathbf{v}) + V_3(\mathbf{x}, \mathbf{v}),$$

defined in the proof of Theorem 1 in (34)-(37). Under the Assumptions 1-3, 5-7, and the conditions on parameters α, β from Theorem 1, we follow its proof, we obtain the recursive step (52) as

$$\mathbb{E}[V(\mathbf{z}(k+1))] \leq (1 - \frac{h_k(2\epsilon_2\epsilon_4 - h_k\eta\epsilon_3\epsilon_5)}{2\epsilon_3\epsilon_4})\mathbb{E}[V(\mathbf{z}(k))] + \frac{\eta h_k^2}{2}(C_1\Delta^2 + C_2\sigma^2).$$

Note: The stepsize condition $0 < h < \min\left\{\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}, \frac{2\epsilon_3}{\epsilon_2}\right\}$ was not necessary for obtaining (52). Unlike the Proof of Theorem 1, we will not show contraction from $k = 0$; since it is enough to show contraction for large enough k to prove Theorem 2. So, we do not need the $0 < h < \min\left\{\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}, \frac{2\epsilon_3}{\epsilon_2}\right\}$ condition. Instead, we will leverage a diminishing stepsize for this purpose.

Let $\bar{C} := \frac{\eta}{2}(C_1\Delta^2 + C_2\sigma^2)$. Substituting $h_k = \frac{\gamma}{k+1}$,

$$\mathbb{E}[V(\mathbf{z}(k+1))] \leq \left(1 - \frac{\frac{\gamma}{k+1}\left(2\epsilon_2\epsilon_4 - \frac{\gamma}{k+1}\eta\epsilon_3\epsilon_5\right)}{2\epsilon_3\epsilon_4}\right)\mathbb{E}[V(\mathbf{z}(k))] + \frac{\gamma^2\bar{C}}{(k+1)^2}.$$

Expanding the coefficient of $\mathbb{E}[V(\mathbf{z}(k))]$,

$$1 - \frac{\gamma}{k+1}\cdot\frac{2\epsilon_2\epsilon_4}{2\epsilon_3\epsilon_4} + \frac{\gamma}{k+1}\cdot\frac{\frac{\gamma}{k+1}\eta\epsilon_3\epsilon_5}{2\epsilon_3\epsilon_4} = 1 - \frac{\gamma\epsilon_2}{\epsilon_3(k+1)} + \frac{\gamma^2\eta\epsilon_5}{2\epsilon_4(k+1)^2}.$$

Thus,

$$\mathbb{E}[V(\mathbf{z}(k+1))] \leq \left(1 - \frac{\gamma\epsilon_2}{\epsilon_3(k+1)} + \frac{\gamma^2\eta\epsilon_5}{2\epsilon_4(k+1)^2}\right)\mathbb{E}[V(\mathbf{z}(k))] + \frac{\gamma^2\bar{C}}{(k+1)^2}$$

For sufficiently large k, since $\frac{1}{(k+1)^2} = o\left(\frac{1}{k+1}\right)$, there exists K such that for all $k \geq K$,

$$\frac{\gamma^2\eta\epsilon_5}{2\epsilon_4(k+1)^2} \leq \frac{1}{2}\cdot\frac{\gamma\epsilon_2}{\epsilon_3(k+1)}.$$

Hence,

$$1 - \frac{\gamma\epsilon_2}{\epsilon_3(k+1)} + \frac{\gamma^2\eta\epsilon_5}{2\epsilon_4(k+1)^2} \leq 1 - \frac{1}{2}\cdot\frac{\gamma\epsilon_2}{\epsilon_3(k+1)}.$$

Defining $\mu = \frac{\gamma\epsilon_2}{2\epsilon_3}$, we obtain for all sufficiently large $k \geq K$,

$$\mathbb{E}[V(\mathbf{z}(k+1))] \leq \left(1 - \frac{\mu}{k+1}\right)\mathbb{E}[V(\mathbf{z}(k))] + \frac{\gamma^2\bar{C}}{(k+1)^2}.$$

We require γ such that $\mu > 1$. Next, we unroll the above from k to K:

$$\mathbb{E}[V(\mathbf{z}(k))] \leq \left(\prod_{i=K+1}^{k}\left(1 - \frac{\mu}{i}\right)\right)\mathbb{E}[V(\mathbf{z}(K))] + \sum_{j=K+1}^{k}\left(\prod_{m=j+1}^{k}\left(1 - \frac{\mu}{m}\right)\right)\frac{\gamma^2\bar{C}}{j^2}.$$

We bound the coefficients as follows. Using $1 - x \leq e^{-x}$,

$$\prod_{i=K+1}^{k}\left(1 - \frac{\mu}{i}\right) \leq \exp\left(-\mu\sum_{i=K+1}^{k}\frac{1}{i}\right).$$

Since

$$\sum_{i=K+1}^{k}\frac{1}{i} \geq \ln(k+1) - \ln(K+1),$$

we obtain

$$\prod_{i=K+1}^{k}\left(1-\frac{\mu}{i}\right)\leq\left(\frac{K+1}{k+1}\right)^{\mu}.$$

Similarly,

$$\prod_{m=j+1}^{k}\left(1-\frac{\mu}{m}\right)\leq\left(\frac{j+1}{k+1}\right)^{\mu}.$$

Substituting the above inequalities for the coefficients,

$$\begin{aligned}\mathbb{E}[V(\mathbf{z}(k))]&\leq\left(\frac{K+1}{k+1}\right)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]+\sum_{j=K+1}^{k}\left(\frac{j+1}{k+1}\right)^{\mu}\frac{\gamma^2\bar{C}}{j^2}\\&=\frac{(K+1)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]}{(k+1)^{\mu}}+\frac{\gamma^2\bar{C}}{(k+1)^{\mu}}\sum_{j=K+1}^{k}\frac{(j+1)^{\mu}}{j^2}.\end{aligned}$$

Since $(j+1)^{\mu}\leq 2^{\mu}j^{\mu}$ for $j\geq 1$, we have the summation term

$$\sum_{j=K+1}^{k}\frac{(j+1)^{\mu}}{j^2}\leq 2^{\mu}\sum_{j=K+1}^{k}j^{\mu-2}.$$

For $\mu>1$,

$$\sum_{j=K+1}^{k}j^{\mu-2}\leq\int_{K+1}^{k+1}x^{\mu-2}dx=\frac{(k+1)^{\mu-1}-(K+1)^{\mu-1}}{\mu-1}.$$

Substituting the summation term with this bound,

$$\begin{aligned}\mathbb{E}[V(\mathbf{z}(k))]&\leq\frac{(K+1)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]}{(k+1)^{\mu}}+\frac{\gamma^2\bar{C}\,2^{\mu}}{(k+1)^{\mu}}\frac{(k+1)^{\mu-1}-(K+1)^{\mu-1}}{\mu-1}\\&\leq\frac{(K+1)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]}{(k+1)^{\mu}}+\frac{\gamma^2\bar{C}\,2^{\mu}}{(k+1)^{\mu}}\frac{(k+1)^{\mu-1}}{\mu-1}\\&=\frac{(K+1)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]}{(k+1)^{\mu}}+\frac{\gamma^2\bar{C}\,2^{\mu}}{(k+1)(\mu-1)}.\end{aligned}$$

Then, from (41) the definition of $\mathbf{x}^0$,

$$\mathbb{E}[\|\mathbf{x}(k)-\mathcal{P}_{\mathcal{X}^*}(\mathbf{x}(k))\|^2]\leq\frac{\mathbb{E}[V(\mathbf{z}(k))]}{\epsilon_4}\leq\frac{(K+1)^{\mu}\mathbb{E}[V(\mathbf{z}(K))]}{(k+1)^{\mu}\,\epsilon_4}+\frac{\gamma^2\bar{C}\,2^{\mu}}{(k+1)(\mu-1)\epsilon_4}.$$

The RHS is $\mathcal{O}\left(\frac{1}{(k+1)^{\mu}}\right)+\mathcal{O}\left(\frac{1}{k+1}\right)$. Given $\mu>1$, the dominant term is of order $O\left(\frac{1}{k+1}\right)$. So, the proof is complete.

A.5 PROOF OF THEOREM 3

We denote the following, which will be used throughout this proof:

$$\boldsymbol{K}:=K_n\otimes\boldsymbol{I}_d,\qquad\boldsymbol{H}:=\frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^{\top}\otimes\boldsymbol{I}_d),\qquad\boldsymbol{Q}:=R\Lambda^{-1}R^{\top}\otimes\boldsymbol{I}_d,$$

$$\bar{v}(k):=\frac{1}{n}(\mathbf{1}_n^{\top}\otimes\boldsymbol{I}_d)\boldsymbol{v}(k),\qquad\bar{\zeta}(k):=\frac{1}{n}(\mathbf{1}_n^{\top}\otimes\boldsymbol{I}_d)\boldsymbol{\zeta}(k),\qquad\bar{\boldsymbol{\zeta}}(k):=\mathbf{1}_n\otimes\bar{\zeta}(k)$$

$$\boldsymbol{g}(k):=\nabla F(\boldsymbol{x}(k)),\qquad\bar{\boldsymbol{g}}(k):=\boldsymbol{H}\boldsymbol{g}(k)$$

$$\boldsymbol{g}^0(k):=\nabla F(\bar{\boldsymbol{x}}(k)),\qquad\bar{\boldsymbol{g}}^0(k):=\boldsymbol{H}\boldsymbol{g}^0(k)=\frac{1}{n}(\mathbf{1}_n\otimes\nabla f(\bar{\boldsymbol{x}}(k))).$$

We again follow a Lyapunov-based technique as formulated below. Define the Lyapunov function

$$V(k) := \sum_{i=1}^{4} V_i(k),$$

where

$$\begin{aligned}
V_1(k) &= \frac{1}{2}\|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}},\\
V_2(k) &= \frac{1}{2}\left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}},\\
V_3(k) &= \boldsymbol{x}(k)^\top \boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right),\\
V_4(k) &= f(\bar{x}(k)) - f^* = F(\bar{\boldsymbol{x}}(k)) - f^*.
\end{aligned}$$

As shown in the paragraph following (37) under Assumption 6, from (q-PDGD) we have

$$\bar{v}(k+1) = \bar{v}(k).$$

So, given the initialization $\sum_{i=1}^n v_i(0) = \mathbf{0}_d$, we have $\bar{v}(k) = 0_d$ for all $k \geq 0$. Upon substituting into (q-PDGD),

$$\bar{\boldsymbol{x}}(k+1) = \bar{\boldsymbol{x}}(k) - h\bar{\boldsymbol{g}}(k) - h\bar{\boldsymbol{\zeta}}(k). \tag{53}$$

Next, we establish necessary inequalities which will be utilized later in the Lyapunov analysis. Since ∇F is Lipschitz-continuous with constant $L_f > 0$ and $\rho(\boldsymbol{H}) = 1$, from above we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] \leq L_f^2\mathbb{E}[\|\boldsymbol{x}(k+1) - \boldsymbol{x}(k)\|^2] &= L_f^2\mathbb{E}[\| - h\bar{\boldsymbol{g}}(k) - h\bar{\boldsymbol{\zeta}}(k)\|^2]\\
&\leq 2L_f^2h^2n\sigma^2 + L_f^2h^2\mathbb{E}[\|\bar{\boldsymbol{g}}(k)\|^2],
\end{aligned} \tag{54}$$

where the last inequality follows from unbiased stochastic gradients under Assumption 7. By definition and Lipschitz continuity from Assumption 2, the following holds:

$$\mathbb{E}[\|\boldsymbol{g}^0(k) - \boldsymbol{g}(k)\|^2] \leq L_f^2\mathbb{E}[\|\bar{\boldsymbol{x}}(k) - \boldsymbol{x}(k)\|^2] = L_f^2\mathbb{E}[\|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}}] = L_f^2\|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}}, \tag{55}$$

$$\mathbb{E}[\|\bar{\boldsymbol{g}}^0(k) - \bar{\boldsymbol{g}}(k)\|^2] = \mathbb{E}[\|\boldsymbol{H}(\boldsymbol{g}^0(k) - \boldsymbol{g}(k))\|^2] \leq \mathbb{E}[\|\boldsymbol{g}^0(k) - \boldsymbol{g}(k)\|^2] \leq L_f^2\|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}}. \tag{56}$$

From Polyak–Łojasiewicz inequality under Assumption 4, the global objective satisfies

$$\|\nabla f(\bar{\boldsymbol{x}}(k))\|^2 \geq 2\nu\big(f(\bar{\boldsymbol{x}}(k)) - f^*\big). \tag{57}$$

Equivalently,

$$\|\bar{\boldsymbol{g}}^0(k)\|^2 = \frac{1}{n}\|\nabla f(\bar{\boldsymbol{x}}(k))\|^2 \geq \frac{2\nu}{n}\big(f(\bar{\boldsymbol{x}}(k)) - f^*\big). \tag{58}$$

Next, we upper bound the first component $\mathbb{E}[V_1(k+1)]$ in the Lyapunov function $V(k+1)$.

$$
\begin{aligned}
&\mathbb{E}[V_1(k+1)] \\
&= \frac{1}{2}\mathbb{E}[\|\boldsymbol{x}(k+1)\|_{\boldsymbol{K}}^2] \\
&= \frac{1}{2}\mathbb{E}[\|\boldsymbol{x}(k) - h(\alpha\boldsymbol{L}\boldsymbol{x}(k) + \alpha\boldsymbol{L}\boldsymbol{\varepsilon}(k) + \beta\boldsymbol{v}(k) + \boldsymbol{g}(k) + \boldsymbol{\zeta}(k))\|_{\boldsymbol{K}}^2] \\
&\le \frac{1}{2}\|\boldsymbol{x}(k) - h(\alpha\boldsymbol{L}\boldsymbol{x}(k) + \beta\boldsymbol{v}(k) + \boldsymbol{g}(k))\|_{\boldsymbol{K}}^2 + h^2\alpha^2\rho^2(\boldsymbol{L})\mathbb{E}[\|\boldsymbol{\varepsilon}(k)\|^2] + h^2\mathbb{E}[\|\boldsymbol{\zeta}(k)\|^2] \\
&\le \frac{1}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 - h\alpha\|\boldsymbol{x}(k)\|_{\boldsymbol{L}}^2 + \frac{h^2\alpha^2}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{L}^2}^2 \\
&\quad + \frac{h^2\beta^2}{2}\left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}(k)\right\|_{\boldsymbol{K}}^2 - h\beta\boldsymbol{x}(k)^\top(\boldsymbol{I}_{nd} - h\alpha\boldsymbol{L})\boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}(k)\right) \\
&+ h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} + h^2n\sigma^2 \\
&= \frac{1}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}(k)\|_{h\alpha\boldsymbol{L} - \frac{h^2\alpha^2}{2}\boldsymbol{L}^2}^2 + \frac{h^2\beta^2}{2}\left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k) + \frac{1}{\beta}\boldsymbol{g}(k) - \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|_{\boldsymbol{K}}^2 \\
&\quad - h\beta\boldsymbol{x}(k)^\top(\boldsymbol{I}_{nd} - h\alpha\boldsymbol{L})\boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k) + \frac{1}{\beta}\boldsymbol{g}(k) - \frac{1}{\beta}\boldsymbol{g}^0(k)\right) + h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} \\
&+ h^2n\sigma^2 \\
&\le \frac{1}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}(k)\|_{h\alpha\boldsymbol{L} - \frac{h^2\alpha^2}{2}\boldsymbol{L}^2}^2 + h^2\beta^2\left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|_{\boldsymbol{K}}^2 + h^2\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 \\
&\quad - h\beta\boldsymbol{x}(k)^\top\boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right) + \frac{h}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 + \frac{h}{2}\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 \\
&\quad + \frac{h^2\alpha^2}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{L}^2}^2 + \frac{h^2\beta^2}{2}\left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|_{\boldsymbol{K}}^2 + \frac{h^2}{2}\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 + h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} \\
&+ h^2n\sigma^2 \\
&= \frac{1}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}(k)\|_{h\alpha\boldsymbol{L} - \frac{h}{2}\boldsymbol{K} - \frac{3h^2\alpha^2}{2}\boldsymbol{L}^2}^2 + \frac{h}{2}(1+3h)\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 \\
&\quad - h\beta\boldsymbol{x}(k)^\top\boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right) + \left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|_{\frac{3h^2\beta^2}{2}\boldsymbol{K}}^2 + h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} + h^2n\sigma^2 \\
&\le \frac{1}{2}\|\boldsymbol{x}(k)\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}(k)\|_{h\alpha\boldsymbol{L} - \frac{h}{2}\boldsymbol{K} - \frac{3h^2\alpha^2}{2}\boldsymbol{L}^2 - \frac{h}{2}(1+3h)L_f^2\boldsymbol{K}}^2 \\
&\quad - h\beta\boldsymbol{x}(k)^\top\boldsymbol{K}\left(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right) + \left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|_{\frac{3h^2\beta^2}{2}\boldsymbol{K}}^2 + h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} + h^2n\sigma^2, \qquad (59)
\end{aligned}
$$

where, the first inequality follows from Cauchy-Schwarz inequality and $\rho(K) = 1$; the second equality is due to (q−PDGD); the third equality follows from the spectral properties in Appendix A.1; and the last inequality follows from (55).

Similarly, we bound the second component $\mathbb{E}[V_2(k+1)]$ next.

$$
\begin{aligned}
&\mathbb{E}[V_2(k+1)] \\
&= \frac{1}{2}\mathbb{E}\left[\|\boldsymbol{v}(k+1)+\frac{1}{\beta}\boldsymbol{g}^0(k+1)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}\right] \\
&= \frac{1}{2}\mathbb{E}\left[\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)+h\beta\boldsymbol{L}\boldsymbol{x}(k)+h\beta\boldsymbol{L}\boldsymbol{\varepsilon}(k)+\frac{1}{\beta}(\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k))\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}\right] \\
&\le \frac{1}{2}\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}+h\boldsymbol{x}(k)^\top(\beta\boldsymbol{K}+\alpha\boldsymbol{L})(\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad+\|\boldsymbol{x}(k)\|^2_{\frac{h^2\beta}{2}(\beta\boldsymbol{L}+\alpha\boldsymbol{L}^2)}+\frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}] \\
&\quad+\frac{1}{\beta}(\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)+h\beta\boldsymbol{L}\boldsymbol{x}(k))^\top\left(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}\right)\mathbb{E}[\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)] \\
&\quad+\frac{h^2\beta^2}{2}\mathbb{E}[\|\boldsymbol{L}\boldsymbol{\varepsilon}(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}]+\mathbb{E}[(h\beta\boldsymbol{L}\boldsymbol{\varepsilon}(\boldsymbol{k}))^\top\left(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}\right)\frac{1}{\beta}\left(\boldsymbol{g}^0(k+1)-\boldsymbol{g}(k)^0\right)] \\
&\le \frac{1}{2}\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}+h\boldsymbol{x}(k)^\top(\beta\boldsymbol{K}+\alpha\boldsymbol{L})(\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad+\|\boldsymbol{x}(k)\|^2_{\frac{h^2\beta}{2}(\beta\boldsymbol{L}+\alpha\boldsymbol{L}^2)}+\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\frac{h}{2\beta}(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K})} \\
&\quad+\frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}]+\frac{1}{2h\beta}\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}] \\
&\quad+\frac{h^2\beta^2}{2}\|\boldsymbol{L}\boldsymbol{x}(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}+\frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}] \\
&\quad+h^2\beta^2\rho\left(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}\right)\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}+\frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}] \\
&\le \frac{1}{2}\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}+h\boldsymbol{x}(k)^\top(\beta\boldsymbol{K}+\alpha\boldsymbol{L})(\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad+\|\boldsymbol{x}(k)\|^2_{h^2\beta(\beta\boldsymbol{L}+\alpha\boldsymbol{L}^2)}+\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\frac{h}{2\beta}(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K})} \\
&\quad+\left(\frac{3}{2\beta^2}+\frac{1}{2h\beta}\right)\mathbb{E}[\|\boldsymbol{g}^0(k+1)-\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}]+h^2\beta^2\left(\frac{1}{\rho_2(L)}+\frac{\alpha}{\beta}\right)\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} \\
&\le \frac{1}{2}\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}+h\boldsymbol{x}(k)^\top(\beta\boldsymbol{K}+\alpha\boldsymbol{L})(\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad+\|\boldsymbol{x}(k)\|^2_{h^2\beta(\beta\boldsymbol{L}+\alpha\boldsymbol{L}^2)}+\|\boldsymbol{v}(k)+\frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\frac{h}{2\beta}(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K})} \\
&\quad+h\left(\frac{3h}{2\beta^2}+\frac{1}{2\beta}\right)\left(\frac{1}{\rho_2(L)}+\frac{\alpha}{\beta}\right)\left(L_f^2\|\bar{\boldsymbol{g}}(k)\|^2+2L_f^2n\sigma^2\right) \\
&+h^2\beta^2\left(\frac{1}{\rho_2(L)}+\frac{\alpha}{\beta}\right)\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}, \qquad (60)
\end{aligned}
$$

where, the first inequality follows from Cauchy-Schwarz inequality; the second equality is due to (q-PDGD); the third equality follows from the spectral properties in Appendix A.1; the third inequality holds since $\rho(Q+\frac{\alpha}{\beta}K)\le\rho(Q)+\frac{\alpha}{\beta}\rho(K)$, the spectral properties in Appendix A.1, $\rho(K)=1$; and the last inequality holds follows from (54).

We bound the third component $\mathbb{E}[V_3(k+1)]$ next.

$$
\begin{aligned}
&\mathbb{E}[V_3(k+1)] \\
&= \mathbb{E}[\boldsymbol{x}(k+1)^\top \boldsymbol{K}(\boldsymbol{v}(k+1) + \frac{1}{\beta}\boldsymbol{g}^0(k+1))] \\
&= \mathbb{E}[(\boldsymbol{x}(k) - h(\alpha \boldsymbol{L}\boldsymbol{x}(k) + \beta \boldsymbol{v}(k) + \boldsymbol{g}(k) + \boldsymbol{g}^0(k) - \boldsymbol{g}^0(k) + \boldsymbol{\zeta}(k) + \alpha \boldsymbol{L}\boldsymbol{\varepsilon}(k)))^\top \boldsymbol{K} \\
&\quad \times (\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k) + h\beta \boldsymbol{L}\boldsymbol{x}(k) + h\beta \boldsymbol{L}\boldsymbol{\varepsilon}(k) + \frac{1}{\beta}(\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)))] \\
&= \boldsymbol{x}(k)^\top (\boldsymbol{K} - h(\alpha + h\beta^2)\boldsymbol{L})(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad + \|\boldsymbol{x}(k)\|^2_{h\beta(\boldsymbol{L} - h\alpha \boldsymbol{L}^2)} + \frac{1}{\beta}\boldsymbol{x}(k)^\top (\boldsymbol{K} - h\alpha \boldsymbol{L})\mathbb{E}[\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)] \\
&\quad - h\beta \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{K}} - h(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k))^\top \boldsymbol{K}\mathbb{E}[\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)] \\
&\quad - h(\boldsymbol{g}(k) - \boldsymbol{g}^0(k))^\top \boldsymbol{K}(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k) + h\beta \boldsymbol{L}\boldsymbol{x}(k) + \frac{1}{\beta}\mathbb{E}[\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)]) \\
&\quad + \frac{1}{\beta}\mathbb{E}[(\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k))^\top \boldsymbol{K}\boldsymbol{x}(k+1)] - h^2\alpha\beta \mathbb{E}[\boldsymbol{\varepsilon}(k)^\top \boldsymbol{L}\boldsymbol{K}\boldsymbol{L}\boldsymbol{\varepsilon}(k)] \\
&- h^2\beta \mathbb{E}[\boldsymbol{\zeta}(k)^\top \boldsymbol{K}\boldsymbol{L}\boldsymbol{\varepsilon}(k)] \\
&\leq \boldsymbol{x}(k)^\top (\boldsymbol{K} - h\alpha \boldsymbol{L})(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)) + \frac{h^2\beta^2}{2}\|\boldsymbol{L}\boldsymbol{x}(k)\|^2 \\
&\quad + \frac{h^2\beta^2}{2}\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{K}} + \|\boldsymbol{x}(k)\|^2_{h\beta(\boldsymbol{L} - h\alpha \boldsymbol{L}^2)} \\
&\quad + \frac{h}{2}\|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}} + \frac{1}{2h\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] + \frac{h^2\alpha^2}{2}\|\boldsymbol{L}\boldsymbol{x}(k)\|^2 \\
&\quad + \frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] - h\beta\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{K}} \\
&\quad + \frac{h^2}{2}\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{K}} + \frac{1}{2}\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] \\
&\quad + \frac{h}{2}\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 + \frac{h}{2}\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\boldsymbol{K}} + \frac{h^2}{2}\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 \\
&\quad + \frac{h^2\beta^2}{2}\|\boldsymbol{L}\boldsymbol{x}(k)\|^2 + \frac{h^2}{2}\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 + \frac{1}{2\beta^2}\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] \\
&+ \frac{h^2\beta}{2}(n\sigma^2 + \rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}) \\
&= \boldsymbol{x}(k)^\top (\boldsymbol{K} - h\alpha \boldsymbol{L})(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)) + \frac{h}{2}(1+2h)\|\boldsymbol{g}(k) - \boldsymbol{g}^0(k)\|^2 \\
&\quad + \|\boldsymbol{x}(k)\|^2_{h(\beta \boldsymbol{L} + \frac{1}{2}\boldsymbol{K}) + h^2(\frac{\alpha^2}{2} - \alpha\beta + \beta^2)\boldsymbol{L}^2} \\
&\quad + (\frac{1}{2h\beta^2} + \frac{1}{\beta^2} + \frac{1}{2})\mathbb{E}[\|\boldsymbol{g}^0(k+1) - \boldsymbol{g}^0(k)\|^2] - \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{h(\beta - \frac{1}{2} - \frac{h}{2} - \frac{h\beta^2}{2})\boldsymbol{K}} \\
&+ \frac{h^2\beta}{2}(n\sigma^2 + \rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}) \\
&\leq \boldsymbol{x}(k)^\top \boldsymbol{K}(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)) - h\alpha \boldsymbol{x}(k)^\top \boldsymbol{L}(\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)) \\
&\quad + \|\boldsymbol{x}(k)\|^2_{h(\beta \boldsymbol{L} + \frac{1}{2}\boldsymbol{K}) + h^2(\frac{\alpha^2}{2} - \alpha\beta + \beta^2)\boldsymbol{L}^2 + \frac{h}{2}(1+2h)L_f^2 \boldsymbol{K}} \\
&\quad + h(\frac{1}{2\beta^2} + \frac{h}{\beta^2} + \frac{h}{2})\left(L_f^2\|\bar{\boldsymbol{g}}(k)\|^2 + 2L_f^2 n\sigma^2\right) - \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{h(\beta - \frac{1}{2} - \frac{h}{2} - \frac{h\beta^2}{2})\boldsymbol{K}} \\
&+ \frac{h^2\beta}{2}(n\sigma^2 + \rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}),
\end{aligned}
\tag{61}
$$

where, the first inequality follows from the Cauchy-Schwarz inequality, the spectral properties in Appendix A.1, and $\rho(K) = 1$; the second equality is due to (q-PDGD); the third equality follows from the spectral properties in Appendix A.1; and the last inequality follows from (54)-(55).

Finally, we bound the fourth component $\mathbb{E}[V_4(k+1)]$ next.

$$
\begin{aligned}
&\mathbb{E}[V_4(k+1)] \\
&= \mathbb{E}[F(\bar{\boldsymbol{x}}(k+1)) - f^*] \\
&= \mathbb{E}[F(\bar{\boldsymbol{x}}(k)) - f^* + F(\bar{\boldsymbol{x}}(k+1)) - F(\bar{\boldsymbol{x}}(k))] \\
&\le F(\bar{\boldsymbol{x}}(k)) - f^* - h\mathbb{E}[\bar{\boldsymbol{g}}(k)^\top \boldsymbol{g}^0(k)] + \frac{h^2 L_f}{2}\mathbb{E}[\|\bar{\boldsymbol{g}}(k)\|^2] + \frac{L_f}{2}\mathbb{E}\left[\left\|-h\bar{\boldsymbol{\zeta}}(k)\right\|^2\right] \\
&= f(\bar{\boldsymbol{x}}(k)) - f^* - \frac{h}{2}\bar{\boldsymbol{g}}(k)^\top(\bar{\boldsymbol{g}}(k) + \bar{\boldsymbol{g}}^0(k) - \bar{\boldsymbol{g}}(k)) \\
&\quad - \frac{h}{2}(\bar{\boldsymbol{g}}(k) - \bar{\boldsymbol{g}}^0(k) + \bar{\boldsymbol{g}}^0(k))^\top \bar{\boldsymbol{g}}^0(k) + \frac{h^2 L_f}{2}\|\bar{\boldsymbol{g}}(k)\|^2 + \frac{L_f}{2}\left(2h^2 n\sigma^2\right) \\
&\le f(\bar{\boldsymbol{x}}(k)) - f^* - \frac{h}{4}\|\bar{\boldsymbol{g}}(k)\|^2 + \frac{h}{4}\|\bar{\boldsymbol{g}}^0(k) - \bar{\boldsymbol{g}}(k)\|^2 - \frac{h}{4}\|\bar{\boldsymbol{g}}^0(k)\|^2 \\
&\quad + \frac{h}{4}\|\bar{\boldsymbol{g}}^0(k) - \bar{\boldsymbol{g}}(k)\|^2 + \frac{h^2 L_f}{2}\|\bar{\boldsymbol{g}}(k)\|^2 + L_f h^2 n\sigma^2 \\
&= f(\bar{\boldsymbol{x}}(k)) - f^* - \frac{h}{4}(1 - 2hL_f)\|\bar{\boldsymbol{g}}(k)\|^2 + \frac{h}{2}\|\bar{\boldsymbol{g}}^0(k) - \bar{\boldsymbol{g}}(k)\|^2 - \frac{h}{4}\|\bar{\boldsymbol{g}}^0(k)\|^2 + L_f h^2 n\sigma^2 \\
&\le f(\bar{\boldsymbol{x}}(k)) - f^* - \frac{h}{4}(1 - 2hL_f)\|\bar{\boldsymbol{g}}(k)\|^2 + \|\boldsymbol{x}(k)\|^2_{\frac{h}{2}L_f^2\boldsymbol{K}} - \frac{h\nu}{2n}(f(\bar{\boldsymbol{x}}(k)) - f^*) + L_f h^2 n\sigma^2, \qquad (62)
\end{aligned}
$$

where the first inequality holds follows from smoothness of F, (2) and (53); the second inequality is due to Cauchy-Schwarz inequality; the third equality holds since $\bar{\boldsymbol{g}}(\boldsymbol{k})^\top \boldsymbol{g}^0(k) = \boldsymbol{g}(k)^\top \boldsymbol{H}\boldsymbol{g}^0(k) = \boldsymbol{g}(k)^\top \boldsymbol{H}\boldsymbol{H}\boldsymbol{g}^0(k) = \bar{\boldsymbol{g}}(\boldsymbol{k})^\top \bar{\boldsymbol{g}}^0(k)$; and the last inequality follows from (56) and (58).

Combining the above components of $\mathbb{E}[V(k+1)]$ in (59)-(62), we have

$$
\begin{aligned}
&\mathbb{E}[V(k+1)] \\
&= \sum_{i=1}^{4}\mathbb{E}[V_i(k+1)] \\
&\le V(k) - \|\boldsymbol{x}(k)\|^2_{h\alpha\boldsymbol{L} - \frac{h}{2}\boldsymbol{K} - \frac{3h^2\alpha^2}{2}\boldsymbol{L}^2 - \frac{h}{2}(1+3h)L_f^2\boldsymbol{K}} \\
&\quad + \|\boldsymbol{x}(k)\|^2_{h^2\beta(\beta\boldsymbol{L} + \alpha\boldsymbol{L}^2)} + \|\boldsymbol{x}(k)\|^2_{h(\beta\boldsymbol{L} + \frac{1}{2}\boldsymbol{K}) + h^2(\frac{\alpha^2}{2} - \alpha\beta + \beta^2)\boldsymbol{L}^2 + \frac{h}{2}(1+2h)L_f^2\boldsymbol{K}} + \|\boldsymbol{x}(k)\|^2_{\frac{h}{2}L_f^2\boldsymbol{K}} \\
&\quad - \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{h(\beta - \frac{1}{2} - \frac{h}{2} - \frac{h\beta^2}{2})\boldsymbol{K} - \frac{h}{2\beta}\boldsymbol{Q} - \frac{h}{2\beta}\frac{\alpha}{\beta}\boldsymbol{K}} + \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\|^2_{\frac{3h^2\beta^2}{2}\boldsymbol{K}} \\
&\quad - \left(\frac{h}{4}(1 - 2hL_f) - h\left(\frac{h}{\beta^2} + \frac{h}{2} + \frac{1}{2\beta^2}\right)L_f^2 - h\left(\frac{3h}{2\beta^2} + \frac{1}{2\beta}\right)\left(\frac{1}{\rho_2(L)} + \frac{\alpha}{\beta}\right)L_f^2\right)\|\bar{\boldsymbol{g}}(k)\|^2 \\
&\quad - \frac{h\nu}{2n}(f(\bar{\boldsymbol{x}}(k)) - f^*) + \mathcal{C}_{noise} \\
&= V(k) - \|\boldsymbol{x}(k)\|^2_{h\boldsymbol{M}_1 - h^2\boldsymbol{M}_2} - \left\|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}^0(k)\right\|^2_{h\boldsymbol{M}_3 - h^2\boldsymbol{M}_4} \\
&\quad - (h\epsilon_3 - h^2\epsilon_4)\|\bar{\boldsymbol{g}}(k)\|^2 - \frac{h\nu}{2n}(f(\bar{\boldsymbol{x}}(k)) - f^*) + C_A, \qquad (63)
\end{aligned}
$$

where the matrices $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4$ are defined by

$$\begin{aligned}
\boldsymbol{M}_1 &= (\alpha-\beta)\boldsymbol{L} - \frac{1}{2}(2+3L_f^2)\boldsymbol{K}, \\
\boldsymbol{M}_2 &= \beta^2\boldsymbol{L} + (2\alpha^2+\beta^2)\boldsymbol{L}^2 + \frac{5}{2}L_f^2\boldsymbol{K}, \\
\boldsymbol{M}_3 &= (\beta - \frac{1}{2} - \frac{\alpha}{2\beta^2})\boldsymbol{K} - \frac{1}{2\beta}\boldsymbol{Q}, \\
\boldsymbol{M}_4 &= (2\beta^2+\frac{1}{2})\boldsymbol{K},
\end{aligned}$$

and C_A collects all variance terms arising from stochastic gradients and quantization as

$$\begin{aligned}
C_A &= h^2\alpha^2\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} + h^2n\sigma^2 + h\left(\frac{3h}{2\beta^2}+\frac{1}{2\beta}\right)\left(\frac{1}{\rho_2(L)}+\frac{\alpha}{\beta}\right)\left(2L_f^2n\sigma^2\right) \\
&\quad + h^2\beta^2\left(\frac{1}{\rho_2(\boldsymbol{L})}+\frac{\alpha}{\beta}\right)\rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4} + L_fh^2n\sigma^2 + h\left(\frac{h}{\beta^2}+\frac{h}{2}+\frac{1}{2\beta^2}\right)\left(2L_f^2n\sigma^2\right) \\
&\quad + \frac{h^2\beta}{2}(n\sigma^2 + \rho^2(\boldsymbol{L})\frac{nd\Delta^2}{4}) \\
&=: h\tilde{C}_A.
\end{aligned} \tag{64}$$

In the following step, we obtain a recursion of $\mathbb{E}[V(k+1)]$ from (63). Here, we utilize the parametric conditions of α, β, h from the theorem statement. From $\beta + \kappa_1 \leq \alpha$ and $\kappa_1 = \frac{1}{\rho_2(\boldsymbol{L})}(4+\frac{3}{2}L_f^2)$, we have

$$(\alpha-\beta)\rho_2(\boldsymbol{L}) - \frac{1}{2}(2+3L_f^2) \geq 1. \tag{65}$$

From $\beta \geq \kappa_4$, we have

$$(\beta - \frac{1}{2} - \frac{\kappa_2}{2\beta}) - \frac{1}{2\beta\rho_2(\boldsymbol{L})} \geq 1. \tag{66}$$

From $\alpha \leq \kappa_2\beta$ and $\beta \geq \kappa_5$, we have

$$\begin{aligned}
\epsilon_3 &= \frac{1}{4} - \frac{1}{2\beta}\left(\frac{1}{\beta}+\frac{1}{\rho_2(\boldsymbol{L})}+\frac{\alpha}{\beta}\right)L_f^2 \\
&\geq \frac{1}{4} - \frac{1}{2\beta}\left(\frac{1}{\beta}+\frac{1}{\rho_2(\boldsymbol{L})}+\kappa_2\right)L_f^2 \geq \frac{1}{8}.
\end{aligned} \tag{67}$$

From (67), and $0 < h < \frac{\epsilon_3}{\epsilon_4}$, we have

$$h\epsilon_3 - h^2\epsilon_4 > 0. \tag{68}$$

From (65), (66), and $\alpha \leq \kappa_2\beta$, we have

$$\begin{aligned}
\boldsymbol{M}_1 &= (\alpha-\beta)\boldsymbol{L} - \frac{1}{2}(2+3L_f^2)\boldsymbol{K} \\
&\geq (\alpha-\beta)\rho_2(\boldsymbol{L})\boldsymbol{K} - \frac{1}{2}(2+3L_f^2)\boldsymbol{K} \geq \boldsymbol{K},
\end{aligned} \tag{69}$$

$$\boldsymbol{M}_2 = \beta^2\boldsymbol{L} + (2\alpha^2+\beta^2)\boldsymbol{L}^2 + \frac{5}{2}L_f^2\boldsymbol{K} \leq \epsilon_2\boldsymbol{K}, \tag{70}$$

$$\begin{aligned}
\boldsymbol{M}_3 &= (\beta - \frac{1}{2} - \frac{\alpha}{2\beta^2})\boldsymbol{K} - \frac{1}{2\beta}\boldsymbol{Q} \\
&\geq (\beta - \frac{1}{2} - \frac{\kappa_2}{2\beta})\boldsymbol{K} - \frac{1}{2\beta\rho_2(\boldsymbol{L})}\boldsymbol{K} \geq \boldsymbol{K},
\end{aligned} \tag{71}$$

$$\boldsymbol{M}_4 = (2\beta^2+\frac{1}{2})\boldsymbol{K} \leq \epsilon_2\boldsymbol{K}. \tag{72}$$

Denote $\hat{V}(k) = \|\boldsymbol{x}(k)\|^2_{\boldsymbol{K}} + \|\boldsymbol{v}(k) + \frac{1}{\beta}\boldsymbol{g}(k)^0\|^2_{\boldsymbol{K}} + f(\bar{x}(k)) - f^*$. Then, upon substituting in (63) from (69)-(72), and with the notation $\epsilon_1 = \min\{1, \frac{\nu}{2n}\}$, we have

$$\mathbb{E}[V(k+1)] \leq V(k) - h(\epsilon_1 - h\epsilon_2)\hat{V}(k) + C_A. \tag{73}$$

From the Cauchy-Schwarz inequality, we further have

$$\epsilon_6 \hat{V}(k) \leq V(k) \leq \epsilon_5 \hat{V}(k), \tag{74}$$

where $\epsilon_6 = \min\{\frac{1}{2\rho(\boldsymbol{L})}, \frac{\alpha-\beta}{2\alpha}\}$. Since $0 < h < \frac{\epsilon_1}{\epsilon_2}$, the coefficient of $\hat{V}(k)$ in (73) is negative.

Substituting (74) in (73) and $C_A = h\tilde{C}_A$, we have

$$\mathbb{E}[V(k+1)] \leq (1 - hc_1)\mathbb{E}[V(k)] + h\tilde{C}_A, \tag{75}$$

where $c_1 := \frac{\epsilon_1 - h\epsilon_2}{\epsilon_5}$. Since $h < \frac{\epsilon_1}{\epsilon_2}$ and $\epsilon_1 <= 2\sqrt{\epsilon_2\epsilon_5}$, the coefficient $(1 - hc_1) \in [0, 1)$. So, we unroll the recursion (75) from k to 0, to obtain

$$\mathbb{E}[V(k)] \leq (1 - hc_1)^k V(0) + \frac{\tilde{C}_A}{c_1}. \tag{76}$$

Denoting $C := \frac{\tilde{C}_A}{c_1}$, as $k \to \infty$, the transient term vanishes (since $0 <= (1 - hc_1) < 1$), and we have

$$\limsup_{k\to\infty} \mathbb{E}[V(k)] \leq C.$$

Hence, using the relation $\epsilon_6\hat{V}(k) \leq V(k) \leq \epsilon_5\hat{V}(k)$, for all $i \in [n]$ we have

$$\begin{aligned}
\limsup_{k\to\infty} \left(\mathbb{E}[\|x_{i,k} - \bar{x}_k\|^2] + \mathbb{E}[f(\bar{x}_k) - f^*]\right) &\leq \limsup_{k\to\infty} \left(\mathbb{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] + \mathbb{E}[f(\bar{x}_k) - f^*]\right) \\
&\leq \limsup_{k\to\infty} \mathbb{E}[\hat{V}_k] \\
&\leq \limsup_{k\to\infty} \frac{1}{\epsilon_6}\mathbb{E}[V_k] \\
&\leq \frac{C}{\epsilon_6}.
\end{aligned} \tag{77}$$

The proof is complete.

B NETWORK SCALING OF η, ϵ_i, AND THE RESIDUAL CONSTANTS A, B

Recall that L is the Laplacian with $\rho(L) = \lambda_{\max}(L)$ and $\rho_2(L) = \lambda_2(L)$. In the fixed step-size regime for Theorem 1, we defined in Section 5:

$$\epsilon_1 := \max\left\{\frac{1}{\nu_1}\Big(\frac{L_f^2}{2\beta} + \beta\,\rho(L)\Big),\ \frac{1}{2}\right\}, \tag{78}$$

$$\epsilon_2 := \min\left\{\frac{\beta}{2},\ \epsilon_1\nu_1\right\}, \tag{79}$$

$$\epsilon_3 := \max\left\{\epsilon_1 + \frac{1}{2},\ \frac{\epsilon_1}{\rho_2(L)} + \frac{\alpha}{2\beta} + \frac{1}{2}\right\}, \tag{80}$$

$$\epsilon_4 := \min\left\{\epsilon_1 - \frac{1}{2},\ \frac{\alpha}{2\beta} - \frac{1}{2}\right\}, \tag{81}$$

$$\epsilon_5 := \max\left\{\beta^2\rho^2(L) + 3\alpha^2\rho^2(L) + 3L_f^2,\ 3\beta^2\right\}, \tag{82}$$

$$\eta := \sqrt{2}\,\max\left\{\frac{2\epsilon_1}{\rho_2(L)} + \alpha + 1,\ 4\epsilon_1 + 1\right\}, \tag{83}$$

and the residual constants:

$$A = \left(\frac{\epsilon_3 \eta h}{2\epsilon_2 \epsilon_4 - h\eta \epsilon_3 \epsilon_5} \right) \frac{nd}{4} \left(2\alpha^2 + \beta^2\right) \rho^2(L), \tag{84}$$

$$B = \left(\frac{\epsilon_3 \eta h}{2\epsilon_2 \epsilon_4 - h\eta \epsilon_3 \epsilon_5} \right) 2n. \tag{85}$$

We treat the problem parameters (L_f, ν) as $\mathcal{O}(1)$ and focus on understanding the influence of network parameters, specifically the network scale n and the graph Laplacian spectral properties $(\rho(L), \rho_2(L))$, on the residual error. Recall from Lemma 2 that the effective RSI constant is given by

$$\nu_1 = \min\left\{ \frac{\nu}{2n}, \alpha\rho_2(L) - \frac{2nL_f^2 + \nu L_f}{\nu} \right\}.$$

The convergence condition in Theorem 1 requires $\alpha > \frac{2nL_f^2 + \nu L_f}{\nu \rho_2(L)}$. Note that the lower bound for α scales as $\mathcal{O}(n/\rho_2(L))$. We select α to be strictly larger than this lower bound while maintaining the same asymptotic order. Specifically, we set $\alpha = \mathcal{O}(n/\rho_2(L))$ with a sufficiently large constant factor such that the second term in the definition of ν_1 scales as $\mathcal{O}(n)$. Since the first term scales as $\mathcal{O}(1/n)$, for sufficiently large n, the minimum is determined by the first term. Consequently, we proceed with the analysis assuming $\nu_1 = \frac{\nu}{2n} = \mathcal{O}(1/n)$, alongside $\beta = \mathcal{O}(1)$ and $\alpha = \mathcal{O}(n/\rho_2(L))$.

From (78) with $\nu_1 = \mathcal{O}(\frac{1}{n})$ and $\beta = \mathcal{O}(1)$,

$$\epsilon_1 = \mathcal{O}\Big(\max\Big\{ \tfrac{1}{\nu_1}(1 + \rho(L)),\ \frac{1}{2} \Big\} \Big),$$

implying that

$$\epsilon_1 = \mathcal{O}\Big(\max\Big\{ n(1 + \rho(L)),\ \frac{1}{2} \Big\} \Big) = \mathcal{O}\big(n\rho(L)\big). \tag{86}$$

Then $\epsilon_1 \nu_1 = \mathcal{O}(1 + \rho(L))$. So (79) implies

$$\epsilon_2 = \mathcal{O}\big(\min\{1,\ 1 + \rho(L)\} \big) = \mathcal{O}(1). \tag{87}$$

Similarly, using the definitions in (80), (83), (82), along with (86), we have the rest of the constants as

$$\epsilon_3 = \mathcal{O}\left(\frac{\epsilon_1}{\rho_2(L)} + \alpha \right) = \mathcal{O}\left(\frac{n\rho(L)}{\rho_2(L)} \right), \tag{88}$$

$$\eta = \mathcal{O}\left(\frac{\epsilon_1}{\rho_2(L)} + \alpha + \epsilon_1 \right) = \mathcal{O}\left(\frac{n\rho(L)}{\rho_2(L)} \right), \tag{89}$$

$$\epsilon_5 = \mathcal{O}\big(\alpha^2 \rho^2(L)\big) = \mathcal{O}\left(\frac{n^2 \rho^2(L)}{\rho_2^2(L)} \right). \tag{90}$$

Consequently, substituting the asymptotic bounds for ϵ_3, ϵ_5, and η yields

$$\eta\epsilon_3 = \mathcal{O}\left(\frac{n^2 \rho^2(L)}{\rho_2^2(L)} \right), \quad \eta\epsilon_3\epsilon_5 = \mathcal{O}\left(n^4 \frac{\rho^4(L)}{\rho_2^4(L)} \right). \tag{91}$$

For $\epsilon_4 = \min\{\epsilon_1 - \frac{1}{2},\ \frac{\alpha}{2\beta} - \frac{1}{2}\}$, with $\alpha = \mathcal{O}(n/\rho_2(L))$, $\beta = \mathcal{O}(1)$, and (86), we have

$$\epsilon_4 = \mathcal{O}\left(\frac{n}{\rho_2(L)} \right). \tag{92}$$

Let R be the stepsize-dependent factor appearing in A and B in (84)-(85):

$$R \ := \ \frac{\epsilon_3 \eta h}{2\epsilon_2 \epsilon_4 - h\eta \epsilon_3 \epsilon_5}.$$

The stability condition in Theorem 1 requires the stepsize to satisfy $h < \min\left\{\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}, \frac{2\epsilon_3}{\epsilon_2}\right\}$. From the above expressions (86)-(92), we observe that

$$\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5} = \mathcal{O}\left(\frac{\rho_2^3(L)}{n^3\rho^4(L)}\right), \; \frac{2\epsilon_3}{\epsilon_2} = \mathcal{O}\left(\frac{n\rho(L)}{\rho_2(L)}\right).$$

So, the minimum of these two quantities is primarily decided by $\frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}$. Having established this, below we will consider two regimes for stepsize h: one where h is small enough and the other where h is near its maximal permissible value for stability of (q−PDGD).

Case-I. Given the rapid growth of $\eta\epsilon_3\epsilon_5$ with respect to the network condition number $\frac{\rho(L)}{\rho_2(L)}$ and network scale n, we proceed with the analysis in the small stepsize regime where $h\eta\epsilon_3\epsilon_5 \ll 2\epsilon_2\epsilon_4$. This ensures that the first-order descent terms in the Lyapunov analysis dominate the higher-order discretization errors (see (52)). Under the small stepsize regime $h\eta\epsilon_3\epsilon_5 \ll 2\epsilon_2\epsilon_4$, the denominator is dominated by $2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5 \approx 2\epsilon_2\epsilon_4$. From (87), $\epsilon_2 = \mathcal{O}(1)$. Substituting these relations into the definition of R, we obtain:

$$R \approx \frac{\epsilon_3\eta h}{2\epsilon_2\epsilon_4} = \mathcal{O}\left(\frac{\epsilon_3\eta h}{\epsilon_4}\right) = \mathcal{O}\Big(h\,\frac{n\rho^2(L)}{\rho_2(L)}\Big). \tag{93}$$

Also, with $\alpha = \mathcal{O}(n/\rho_2(L))$ and $\beta = \mathcal{O}(1)$,

$$(2\alpha^2 + \beta^2) = \mathcal{O}\Big(\frac{n^2}{\rho_2^2(L)}\Big). \tag{94}$$

Substituting (93)–(94) into definitions of A, B in (84)-(85) yields

$$A = \mathcal{O}\Big(\Big(h\,\frac{n\rho^2(L)}{\rho_2(L)}\Big)\cdot nd\cdot\Big(\frac{n^2}{\rho_2^2(L)}\Big)\cdot\rho^2(L)\Big) = \mathcal{O}\Big(h\,d\,\frac{n^4\rho^4(L)}{\rho_2^3(L)}\Big), \tag{95}$$

$$B = \mathcal{O}\Big(\Big(h\,\frac{n\rho^2(L)}{\rho_2(L)}\Big)\cdot n\Big) = \mathcal{O}\Big(h\,\frac{n^2\rho^2(L)}{\rho_2(L)}\Big). \tag{96}$$

Case-II. Now we consider the maximal stable stepsize regime, i.e., $h = \mathcal{O}\Big(\frac{\epsilon_2\,\epsilon_4}{\eta\,\epsilon_3\,\epsilon_5}\Big)$. Upon substituting from (91),

$$h = \mathcal{O}\Big(\frac{\epsilon_2\,\epsilon_4}{\eta\,\epsilon_3\,\epsilon_5}\Big) = \mathcal{O}\Big(\frac{\rho_2^3(L)}{n^3\rho^4(L)}\Big), \tag{97}$$

Finally, substituting from (97) into the definitions (95)-(96) gives

$$A = \mathcal{O}\Big(\Big(\frac{\rho_2^3(L)}{n^3\rho^4(L)}\Big)\cdot d\,\frac{n^4\rho^4(L)}{\rho_2^3(L)}\Big) = \mathcal{O}\Big(\frac{d\,n}{1}\Big) = \mathcal{O}(d\,n)\,, \tag{98}$$

$$B = \mathcal{O}\Big(\Big(\frac{\rho_2^3(L)}{n^3\rho^4(L)}\Big)\cdot\frac{n^2\rho^2(L)}{\rho_2(L)}\Big) = \mathcal{O}\Big(\frac{\rho_2^2(L)}{n\,\rho^2(L)}\Big). \tag{99}$$

Equations (95)-(96) and (98)-(99) are exactly the scalings quoted in Remark 2.

C ADDITIONAL PLOTS

Under L_f-smoothness (and, when applicable, the PL inequality), the objective suboptimality $\mathbb{E}[F(\mathbf{x})-F(\mathbf{x}^*)]$ is proportional to the squared distance to the solution $\mathbb{E}[\|\mathbf{x}-\mathbf{x}^*\|^2]$. Consequently, the linear dependence on the noise variances Δ^2 and σ^2 predicted by our bounds manifests directly in the steady-state objective gaps.

In particular, from Theorem 1, the steady-state neighborhood radius under a constant step-size is limited by the quantization resolution Δ and the stochastic gradient variance σ^2. Figure 4 explicitly demonstrates the sensitivity of the asymptotic objective gap to these parameters across varying network sizes. The results exhibit a strict linear dependence on both the squared quantization step Δ^2 and the noise variance σ^2. This empirical linear trend directly validates the derived theoretical bound from Theorem 1, which bounds the error proportionally to $A\Delta^2 + B\sigma^2$. Furthermore, Figure 4 captures the explicit

network scaling effects established in Remark 2. In the left subfigure, the slope of the error with respect to Δ^2 increases as the number of agents n grows (with σ^2 fixed at 0.5). This aligns perfectly with the theoretical scaling of A, which depends linearly on the number of agents n. Conversely, in the right subfigure, the slope of the error with respect to σ^2 decreases as n increases (with number of communication bits fixed at 14). This clearly corroborates the derived theoretical behavior where B is inversely proportional to n, demonstrating the variance-reduction benefit of larger networks.

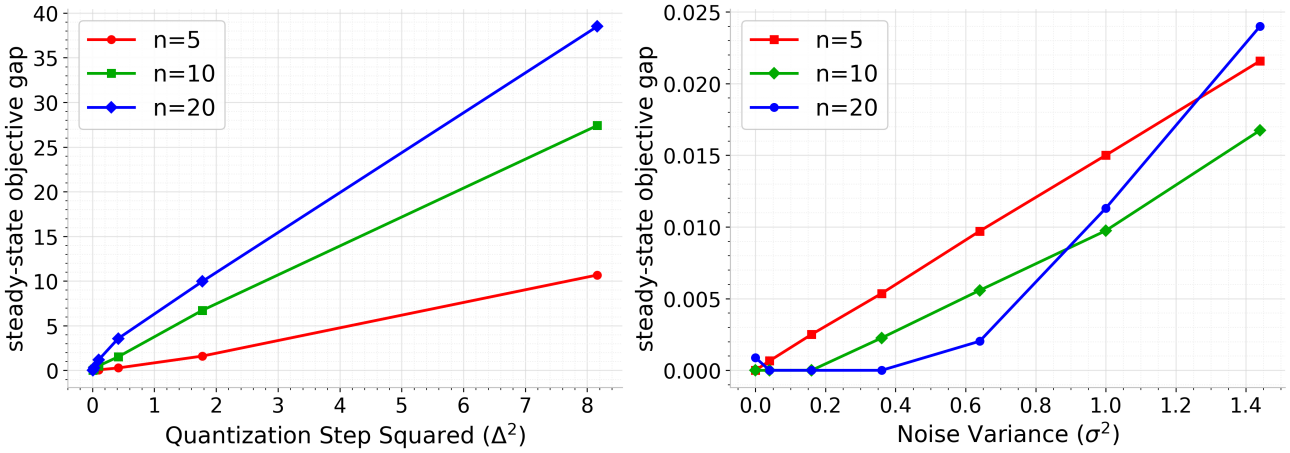

Figure 4: Asymptotic objective gap of `q-PDGD` for Example 1 versus (left) quantization step squared Δ^2 and (right) gradient-noise variance σ^2, for multiple network sizes n. The linear trends corroborate the $\mathcal{O}(A\Delta^2 + B\sigma^2)$ steady-state bound from Theorem 1 and the network scaling discussed in Remark 2.

D EMPIRICAL EVALUATION ON DISTRIBUTED DEEP LEARNING

The theoretical results in Section A are stated under RSI (Assumption 3) or PL (Assumption 4) on the global objective. Modern deep networks satisfy neither condition globally: the loss landscape is non-convex with many saddle points and local minima. Nevertheless, a practical algorithm motivated by these analyses should remain robust in this regime, since (i) the consensus and dual-tracking mechanism is geometry-agnostic and (ii) PL-like behavior is widely observed locally near minima of overparameterized networks. We therefore stress-test `q-PDGD` against representative quantized decentralized baselines on two image-classification benchmarks where our assumptions are formally violated, and show that the consensus advantage predicted by Theorem 2 continues to hold.

Setup. We train a small convolutional network collaboratively across $n = 100$ agents arranged in a ring topology, communicating once per local SGD step over a 4-bit stochastic uniform quantizer on the range $[-1, 1]$ (Eq. (1)). The model is a two-layer CNN ($1\to8\to16$ channels with 5×5 kernels and 2×2 max-pooling, followed by a linear classifier) totaling $\approx 11{,}000$ parameters per agent. We evaluate on **MNIST** and **Fashion-MNIST**; the training set is shuffled and split IID across the 100 agents (600 images each), so all heterogeneity in the trajectories arises from local stochastic gradients and quantization noise rather than from data heterogeneity. Each agent draws a minibatch of size 32 and runs for $T = 500$ communication rounds. All agents are initialized identically, and the random seed (controlling minibatches and quantization noise) is reset before each algorithm, so any difference between curves is attributable solely to the consensus mechanism.

Baselines and hyperparameters. We compare `q-PDGD` against `q-DGD` [Yuan et al., 2016], `q-DCSG` [Dutta and Doan, 2024], and `CHOCO-SGD` [Koloskova* et al., 2020]. The mixing matrix is $W = I - \omega L$ with $\omega = 1/(\Delta_{\max} + 1)$, where $\Delta_{\max}$ is the maximum degree. For a fair comparison the gradient stepsize is fixed to $h = 0.05$ across all algorithms, so the comparison isolates the consensus dynamics. `q-PDGD` uses $\alpha = 0.5$, $\beta = 0.5$; `q-DCSG` uses diminishing $\alpha_k = 0.05/(1 + 0.005\,k)$, $\beta_k = 0.5/(1 + 0.005\,k)$; `CHOCO-SGD` uses consensus stepsize $\gamma = 0.5$.

Metrics. At each evaluation round we record (i) per-agent test accuracy $\{\mathrm{acc}(x_i)\}_{i=1}^n$, (ii) the consensus error $\frac{1}{n}\sum_i \|x_i - \bar{x}\|$, (iii) mean training loss across agents, and (iv) the test accuracy of the *averaged-parameter* model $\bar{x} = \frac{1}{n}\sum_i x_i$, which is the model that would actually be deployed.

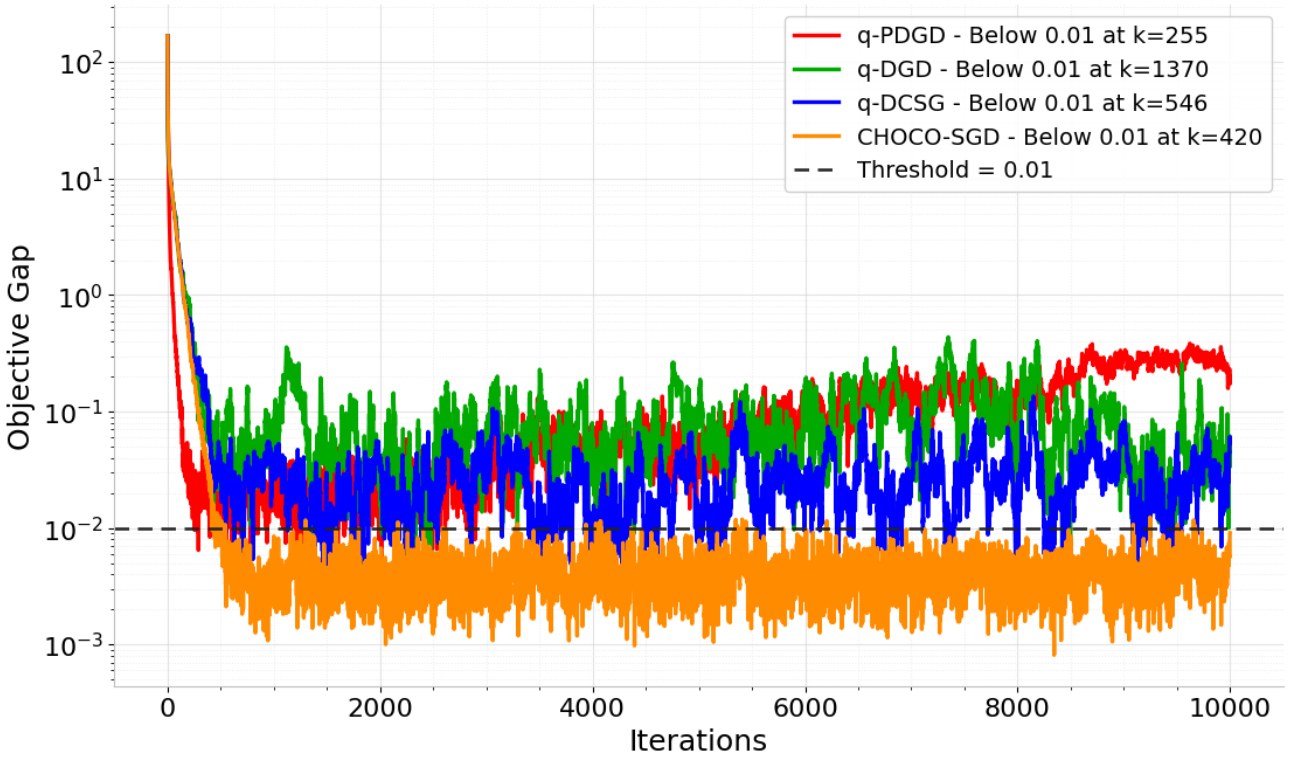

Figure 5: Iterations to reach a target objective threshold: q−PDGD versus q−DGD and DCSG on distributed least squares. q−PDGD requires markedly fewer iterations to reach the target threshold.

Results. Figures 6–9 report the comparison. Three observations are consistent across both datasets and align with our theory.

(1) Smallest consensus residual. Panel (b) of Figs. 6 and 8 (log scale) shows that q−PDGD stabilizes at a consensus error roughly $4-8\times$ below CHOCO−SGD and an order of magnitude below q−DGD/q−DCSG. This is the structural advantage predicted by the dual variable v_k in Eq. (q−PDGD): it accumulates and feeds back disagreement induced by quantized mixing, suppressing the residual neighborhood radius even when the quantizer is aggressive (4 bits).

(2) Best deployed (consensus) model. Panel (d) shows that the averaged model from q−PDGD dominates throughout training. Tighter consensus means $\bar{x}$ remains close to every x_i, so averaging does not destroy useful local progress; on Fashion-MNIST, q−DGD loses several accuracy points purely to disagreement.

(3) Tight cross-agent distribution. Figures 7 and 9 report the spread of final per-agent accuracies. On MNIST, q−PDGD reaches $\mu = 0.966$, $\sigma = 0.0050$, against q−DGD's $\mu = 0.906$, $\sigma = 0.0277$; on Fashion-MNIST, $\mu = 0.819$, $\sigma = 0.0158$ against $\mu = 0.749$, $\sigma = 0.0205$. The histograms are visibly concentrated for q−PDGD, indicating that no agent is left behind.

Although the global-geometry assumptions of Section 4.1 do not hold for these non-convex CNN objectives, the qualitative behavior predicted by our analysis—in particular, a markedly smaller asymptotic consensus neighborhood than competing quantized decentralized methods—persists, suggesting that the primal–dual disagreement-tracking mechanism is useful well beyond the regime covered by the formal guarantees.

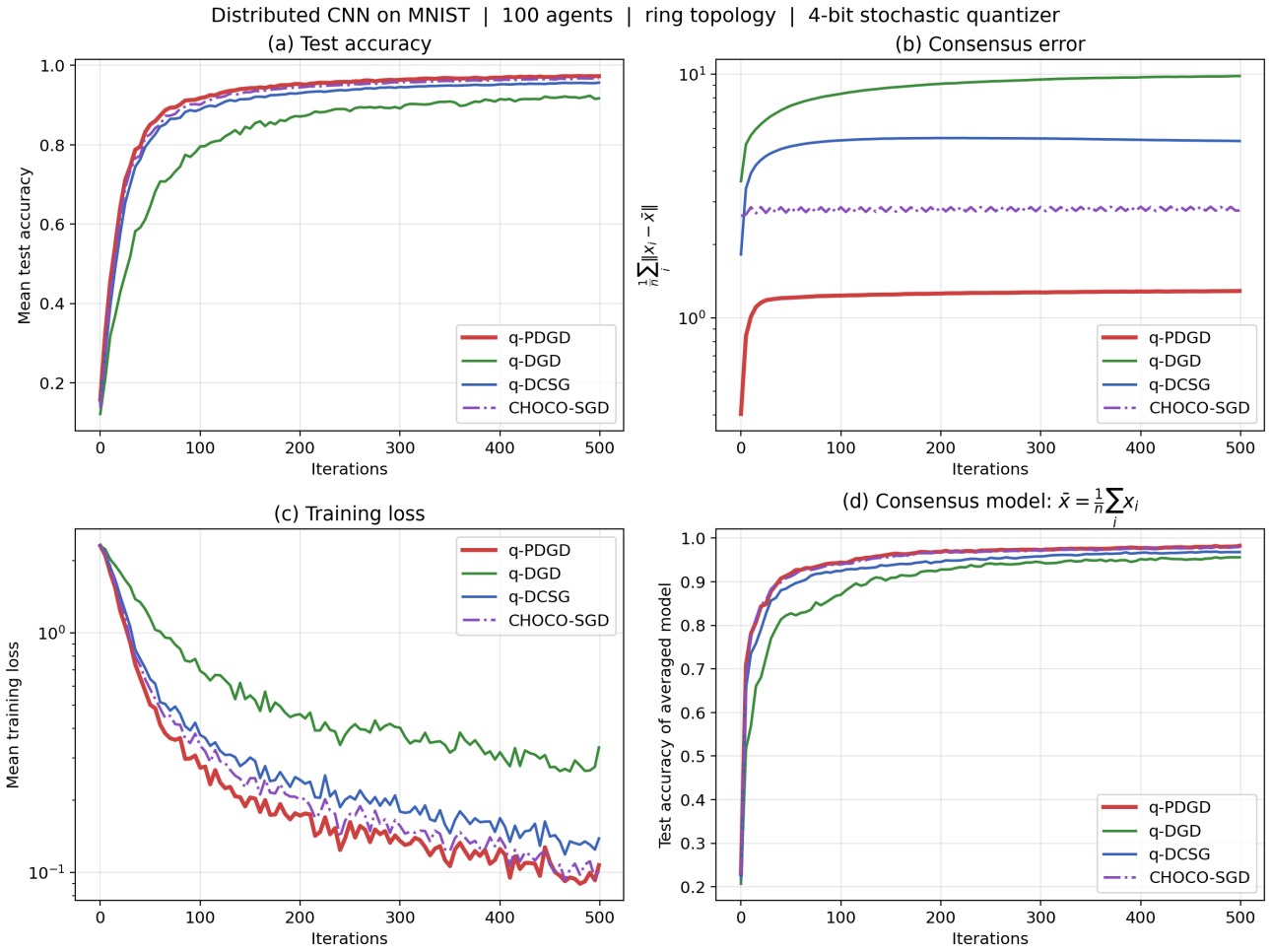

Figure 6: **MNIST,** 100 **agents, ring topology,** 4**-bit stochastic quantizer.** (a) Mean per-agent test accuracy. (b) Consensus error $\frac{1}{n}\sum_i \|x_i - \bar{x}\|$ (log scale): `q-PDGD` is roughly an order of magnitude lower than the baselines. (c) Mean training loss. (d) Test accuracy of the averaged (deployed) model $\bar{x} = \frac{1}{n}\sum_i x_i$.

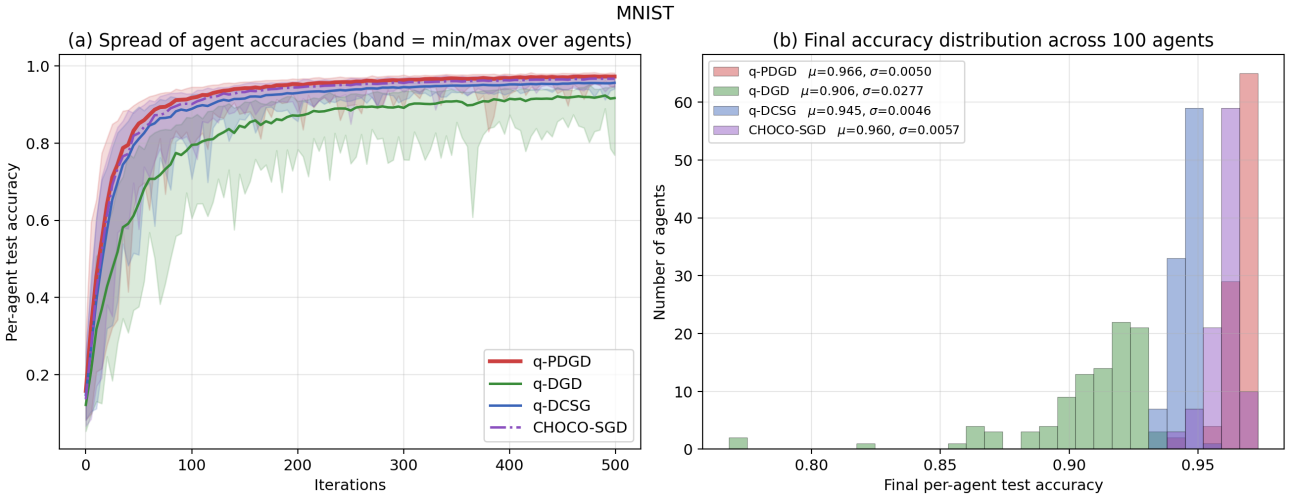

Figure 7: **MNIST: dispersion across the** 100 **agents.** (a) Mean accuracy with shaded min/max envelope. (b) Histogram of final per-agent test accuracies. `q-PDGD` produces the right-most and most concentrated distribution.

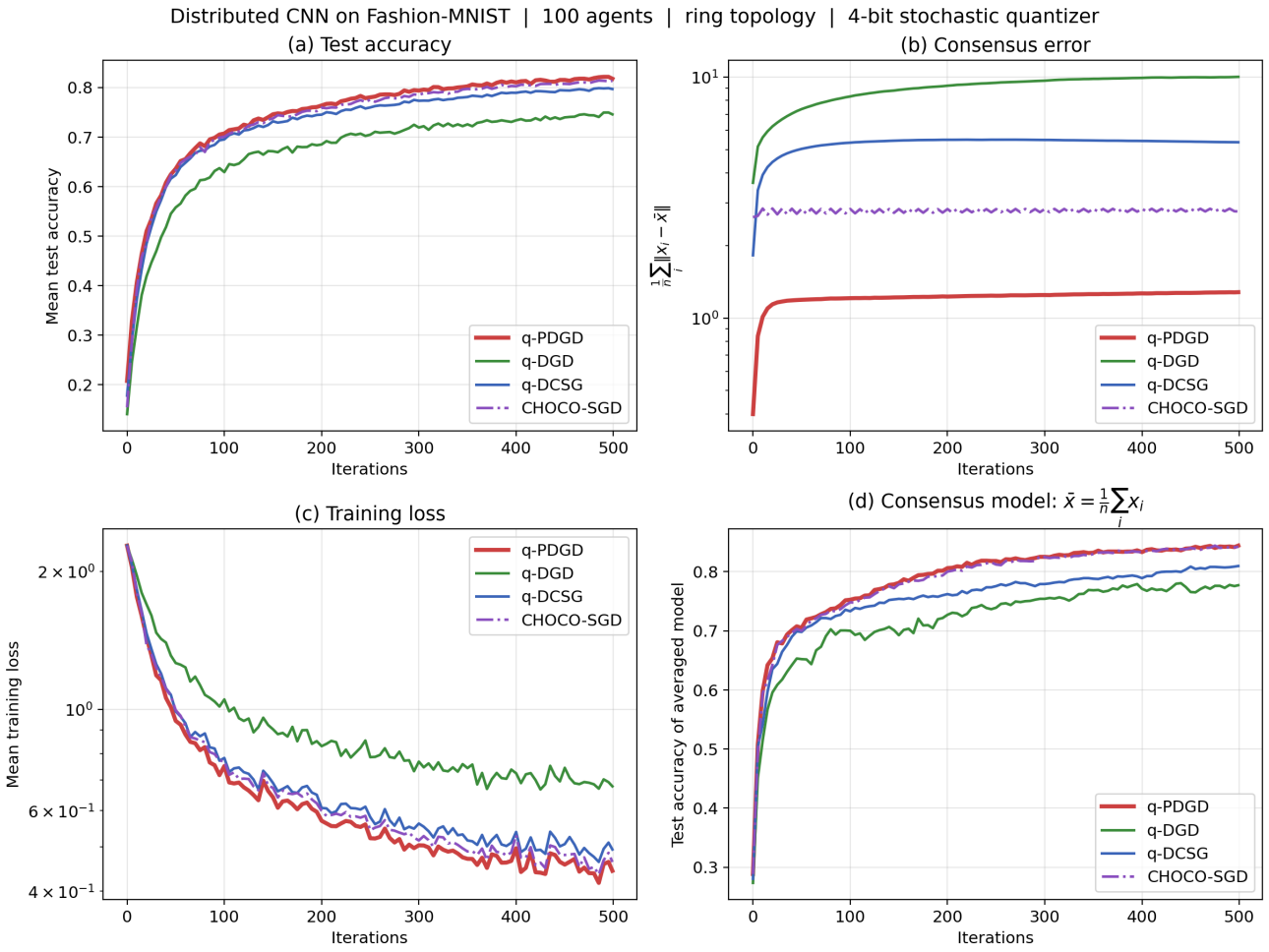

Figure 8: **Fashion-MNIST,** 100 **agents, ring topology,** 4**-bit stochastic quantizer.** Same panels as Fig. 6. The harder task amplifies the gap between algorithms in panel (d): tight consensus directly translates into a better deployed model.

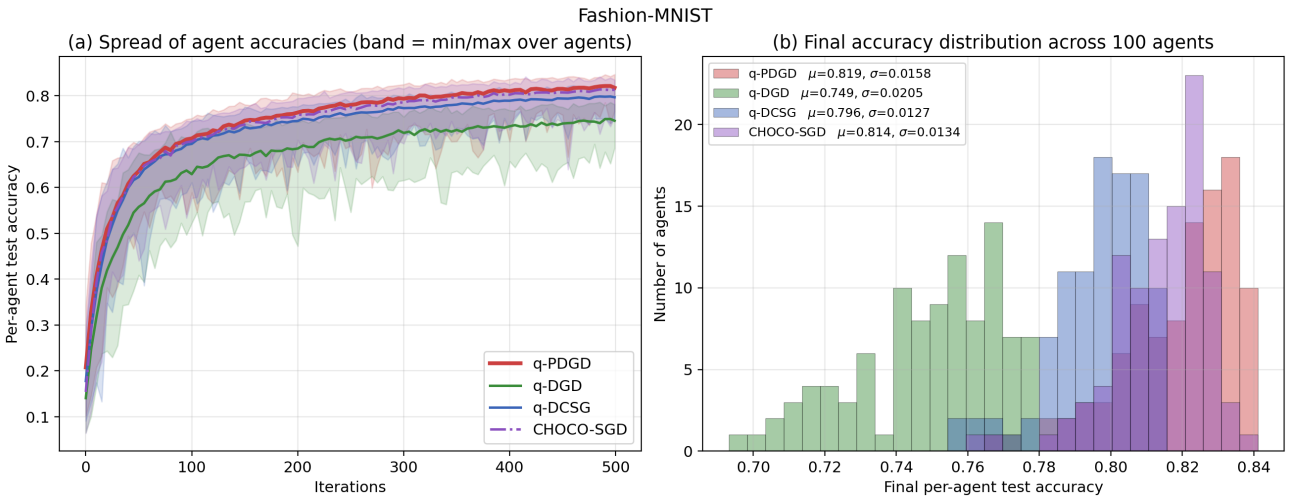

Figure 9: **Fashion-MNIST: dispersion across the** 100 **agents.** Final per-agent accuracies: `q-PDGD` μ=0.819, σ=0.0158; `CHOCO-SGD` μ=0.814, σ=0.0134; `q-DCSG` μ=0.796, σ=0.0127; `q-DGD` μ=0.749, σ=0.0205.