Scalable Average Consensus with Compressed Communications

Mohammad Taha Toghani, César A. Uribe
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005
{mttoghani,cauribe}@rice.edu

Abstract

We propose a new decentralized average consensus algorithm with compressed communication that scales linearly with the network size n. We prove that the proposed method converges to the average of the initial values held locally by the agents of a network when agents are allowed to communicate with compressed messages. The proposed algorithm works for a broad class of compression operators (possibly biased), where agents interact over arbitrary static, undirected, and connected networks. We further present numerical experiments that confirm our theoretical results and illustrate the scalability and communication efficiency of our algorithm.

1 Introduction

We consider the problem of decentralized average consensus over a network of n agents, where each agent $i \in [n]$ starting from an initial vector $\mathbf{x}_i \in \mathbb{R}^d$, seeks to reach consensus on the global average through communication with its neighbors. Formally, the agents attempt to collaboratively solve the following optimization problem:

$$\mathbf{x}^\star := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{x}-\mathbf{x}_i\|^2, \tag{1}$$

only by sharing information with their local neighbors on the corresponding communication network.

The average consensus problem is at the core of many decentralized problems like inference [1] and optimization [2, 3] which themselves are motivated by a wide range of applications such as decentralized federated learning [4], distributed localization and tracking [5], distributed sensor fusion [6], distributed time synchronization [7], etc. These algorithms generally enjoy advantages like parallel computation, privacy, and resiliency to the central party's failure [8]. However, they raise several important challenges such as the existence of adversaries, connection failure, synchronization, communication overhead, and scalability [9].

In gossip type algorithms, each node $i \in [n]$ builds a sequence $\{\mathbf{x}_i(t)\}_{t\geq 0}$ over the course of time, by interacting with its neighbors [10–12]. Given a set of initial parameters $\mathbf{x}_i(0)$, for all $i \in [n]$, their objective is to solve (1), i.e., reach consensus on $\overline{\mathbf{x}} := \big(\sum_{i=1}^{n}\mathbf{x}_i(0)\big)/n$. The convergence rate of such algorithms essentially depends on the connectivity of the network over which the agents interact [10].

Decentralized consensus frameworks classically require agents to share their current estimates of the average value with their neighbors. This imposes a significant communication overhead on the

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

network when d, the estimates' dimension, is large [10, 11]. To address this issue, several average consensus methods have been proposed under quantized communication techniques [13–19], wherein the agents reduce the number of transmitted bits per communication round. However, convergence is generally not exact (i.e., only to some point close true average), or increasingly finer quantization is required. Recently, the authors in [20, 21] used an error-feedback scheme to provide algorithms with exact consensus to the average. Nevertheless, the dependency on the network topology and the number of agents is suboptimal. Recent studies have explored these phenomena in optimization and inference problems [22–25].

The network size n plays an essential role in the scalability of the gossip-type algorithms. Network structures with low connectivity, e.g., path and ring, have quadratic mixing times $\mathcal{O}(n^2)$ [26], i.e., the number of iterations necessary for them to reach consensus grows quadratically with n. In [12], the author suggested a momentum-based approach that implicitly improves the dependence of mixing time by a factor n. This technique has been extended to optimization and social learning problems [1, 27].

In this paper, we jointly address the (i) *communication-efficiency* and (ii) *scalability* challenges for the decentralized average consensus problem. Motivated by [12, 21], we propose a scalable algorithm that requires agents to communicate compressed messages using a class of randomized compression operators. Prior efforts have proposed either scalable [12] or communication-efficient [21] algorithms, while our work exploits both.

Our contributions can be summarized as follows:

- We present a novel scalable and communication-efficient algorithm for the average consensus problem in a decentralized setup.
- Under an appropriate compression operator, we provide convergence guarantees for our proposed algorithm as well as an extension of the algorithm in [12]. Moreover, we show the convergence rate depends linearly on the number of nodes.
- We present the communication advantages of our algorithm through numerical results on two classes of networks with low connectivity.

The remainder of this paper is structured as follows. In Section 2, we describe our problem setup and propose our algorithm, *Scalable Compressed Gossip*. In Section 3, we state our theoretical results and the convergence proofs. Section 4 provides numerical results for the proposed algorithm. Finally, conclusions and future works are remarked in Section 5.

$\diamond$ **Notation:** We write $[n]$ to denote the set $\{1, \dots, n\}$. We use the bolding notation for vectors and matrices. For a matrix $\mathbf{X}$, we write $\mathbf{X}_{ij}$ to denote the entry in the i-th row and j-th column. We use $\mathbf{I}_n$ for the identity matrix of size $n \times n$ as well as $\mathbf{1}_n$ for the vector of all one with size n, where we may drop n for brevity. We refer to agents by subscripts. We write $\lambda_i(\mathbf{W})$ to denote the i-th largest eigenvalue of matrix $\mathbf{W}$, in terms of magnitude. We denote $\|\mathbf{x}\|$ and $\|\mathbf{X}\|_F$ respectively as 2-norm of vector $\mathbf{x}$ and Frobenius norm of matrix $\mathbf{X}$. We refer to matrix norm of a square matrix $\mathbf{W}$ as $\|\mathbf{W}\|$. We denote $\mathbf{A} \otimes \mathbf{B}$ as the Kronecker product of any two matrices $\mathbf{A}$ and $\mathbf{B}$. We write $\mathbf{x}(t)$ in reference to the value of parameter $\mathbf{x}$ at time t.

2 Problem Setup & Algorithm

This section first states the communication setup and describes the class of compression operators used by the proposed algorithm. We then present our scalable and communication-efficient algorithm.

$\diamond$ **Communication Network:** Consider a set of n agents interacting over a fixed, undirected, and connected communication network $\mathcal{G} = \{[n], \mathcal{E}\}$, where $\mathcal{E} \subseteq [n] \times [n]$ is the set of edges. If there is a link between any two agents i and j, then they may exchange information with each other. We denote $\mathcal{N}_i$ as the set of agent i's neighbors as well as $\mathcal{N}_i' = \mathcal{N}_i \cup \{i\}$, for all $i \in [n]$. We denote matrix $\mathbf{W} \in [0, 1]^{n \times n}$ with positive diagonal entries, a proper *mixing matrix* corresponding to network $\mathcal{G}$, if it is symmetric ($\mathbf{W} = \mathbf{W}^\top$), doubly stochastic ($\mathbf{W}\mathbf{1} = \mathbf{W}^\top \mathbf{1} = \mathbf{1}$), and $\mathbf{W}_{ij} = 0$ for $(i, j) \notin \mathcal{E}$, $i \neq j$. We also denote $\delta(\mathbf{W})$ as the *spectral gap* of matrix $\mathbf{W}$, i.e., the gap between the first and second largest eigenvalues of $\mathbf{W}$, which lies in $(0, 1]$. Furthermore, given an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, we define its associated *Metropolis-Hasting* mixing matrix $\mathbf{W} = \mathcal{MH}(\mathcal{G})$ [28]

as follows:

$$\mathbf{W}_{ij} = \begin{cases} \frac{1}{\max\left\{|\mathcal{N}_i'|,|\mathcal{N}_j'|\right\}}, & \text{if } (i,j) \in \mathcal{E} \\ 1 - \sum_{j \neq i} \mathbf{W}_{ij}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$\diamond$ **Compression Operator:** Here, we introduce a class of compression operators that has been widely studied for distributed optimization [21, 29, 30]. We assume the compression operator $Q : \mathbb{R}^d \times \mathcal{Z} \times [0,1) \to \mathbb{R}^d$ satisfies

$$\mathbb{E}_{\boldsymbol{\zeta}} \left\| Q(\mathbf{x}, \boldsymbol{\zeta}, \omega) - \mathbf{x} \right\|^2 \leq \omega^2 \left\| \mathbf{x} \right\|^2, \qquad \forall \mathbf{x} \in \mathbb{R}^d, \tag{3}$$

where $\omega \in [0,1)$, $\boldsymbol{\zeta}$ is a random variable with output space $\mathcal{Z}$, and $\mathbb{E}_{\boldsymbol{\zeta}}[.]$ indicates the expectation over the internal randomness of Q. Note that in (3), $\omega = 0$ implies no compression (i.e., exact communications). Hereafter, we drop $\boldsymbol{\zeta}, \omega$ from Q and $\mathbb{E}$ for simplicity of notation.

The class of randomized operators introduced in (3) embraces a wide range of functions, both sparsification, and quantization, some of which we mention in Example 1 [24].

Example 1. *The following operators fulfill (3):*

- rand_k*: Select k out of d coordinates randomly and mask the rest to zero, $\omega^2 = 1 - k/d$.*
- top_k*: Select k out of d coordinates with highest magnitude and mask the rest to zero, $\omega^2 = 1 - k/d$.*
- qsgd_k*: Round each coordinate of $|\mathbf{x}|/\|\mathbf{x}\|$ to one of the $u = 2^{k-1} - 1$ quantization levels ($k-1$ bits), and one bit for the sign of the coordinate, i.e.,*

$$\mathrm{qsgd}_k(\mathbf{x}) = \frac{\mathrm{sign}(\mathbf{x}).\|\mathbf{x}\|}{u\tau} \left\lfloor u \frac{|\mathbf{x}|}{\|\mathbf{x}\|} + \boldsymbol{\zeta} \right\rfloor, \qquad \boldsymbol{\zeta} \sim [0,1]^d,$$

where $\tau = 1 + \min\left\{d/u^2, \sqrt{d}/u\right\}$, and $\omega^2 = 1 - \tau^{-1}$.

We next propose our method and discuss its features.

$\diamond$ **Algorithm:** We here present our communication-efficient and scalable gossip type algorithm. As we discussed in Section 1, let $\mathbf{x}_i(t)$ be the vector belonging to agent i at time t, for all $i \in [n]$ and $t \geq 0$. Similar to [21], we consider an error-feedback framework, wherein each agent i gradually estimates $\hat{\mathbf{x}}_j(t)$, an approximation of its neighbors' parameters $\mathbf{x}_j(t)$ (including itself), for all $j \in \mathcal{N}_i'$. Algorithm 1 presents a detailed pseudo-code for our method. Each agent i begins with an initial $\mathbf{x}_i(0)$ and a slack parameter $\mathbf{y}_i(0) = \mathbf{x}_i(0)$, besides $\hat{\mathbf{x}}_j(0) = \mathbf{0}$. Lines 3-7 of Algorithm 1 describe the operations for each round of the algorithm. In a nutshell, agent i at round t, (i) computes a compressed version $\mathbf{q}_i(t)$ of the difference between $\mathbf{x}_i(t)$ and $\hat{\mathbf{x}}_i(t)$, (ii) exchanges compressed vectors $\mathbf{q}_i(t)$ and $\mathbf{q}_j(t)$ with each neighbor $j \in \mathcal{N}_i$, (iii) uses $\mathbf{q}_j(t)$ to update $\hat{\mathbf{x}}_j(t+1)$, for all $j \in \mathcal{N}_i'$, then (iv) updates $\mathbf{y}_i(t+1)$ based on $\mathbf{x}_i(t)$ and $\hat{\mathbf{x}}_j(t+1)$, for all $j \in \mathcal{N}_i'$, and finally (v) extrapolates $\mathbf{x}_i(t+1)$ based on $\mathbf{y}_i(t+1)$ and $\mathbf{y}_i(t)$.

Algorithm 1 Scalable Compressed Gossip (*SCG*)

input: initial parameters $\mathbf{x}_i(0) \in \mathbb{R}^d$, for all $i \in [n]$, network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mixing matrix $\mathbf{W}$, stepsize $\gamma \in (0,1]$, operator Q with $\omega \in [0,1)$, momentum $\sigma \in [0,1)$.

1: $\hat{\mathbf{x}}_i(0) := \mathbf{0}$, $\mathbf{y}_i(0) := \mathbf{x}_i(0)$, $\quad \forall i \in [n]$
2: **for** t **in** $0, \ldots, T-1$, in parallel $\forall i \in [n]$ **do**
3: $\quad \mathbf{q}_i(t) := Q(\mathbf{x}_i(t) - \hat{\mathbf{x}}_i(t))$
4: $\quad$ Send $\mathbf{q}_i(t)$ and receive $\mathbf{q}_j(t)$, $\quad$ for all $j \in \mathcal{N}_i$
5: $\quad \hat{\mathbf{x}}_j(t+1) := \hat{\mathbf{x}}_j(t) + \mathbf{q}_j(t)$, $\quad$ for all $j \in \mathcal{N}_i'$
6: $\quad \mathbf{y}_i(t+1) := \mathbf{x}_i(t) + \gamma \sum_{j \in \mathcal{N}_i'} \mathbf{W}_{ij} \left(\hat{\mathbf{x}}_j(t+1) - \hat{\mathbf{x}}_i(t+1)\right)$
7: $\quad \mathbf{x}_i(t+1) := (1+\sigma)\, \mathbf{y}_i(t+1) - \sigma\, \mathbf{y}_i(t)$
8: **end for**

We now state a matrix notation for our algorithm. Let matrices $\mathbf{X}(t) = \big[\mathbf{x}_1(t), \dots, \mathbf{x}_n(t)\big]^\top$, $\hat{\mathbf{X}}(t) = \big[\hat{\mathbf{x}}_1(t), \dots, \hat{\mathbf{x}}_n(t)\big]^\top$, $Q(\mathbf{X}) = \big[Q(\mathbf{x}_1), \dots, Q(\mathbf{x}_n)\big]^\top$, $\overline{\mathbf{X}} = \big[\overline{\mathbf{x}}, \dots, \overline{\mathbf{x}}\big]^\top$, as well as $\mathbf{Y}(t) = \big[\mathbf{y}_1(t), \dots, \mathbf{y}_n(t)\big]^\top$ with size $n \times d$ be the concatenation of their corresponding vectors. Then, Algorithm 1 may be written as follows:

$$\begin{aligned} \hat{\mathbf{X}}(t+1) &\coloneqq \hat{\mathbf{X}}(t) + Q\big(\mathbf{X}(t) - \hat{\mathbf{X}}(t)\big), \\ \mathbf{Y}(t+1) &\coloneqq \mathbf{X}(t) + \gamma\big(\mathbf{W} - \mathbf{I}\big)\hat{\mathbf{X}}(t+1), \\ \mathbf{X}(t+1) &\coloneqq (1+\sigma)\mathbf{Y}(t+1) - \sigma\mathbf{Y}(t), \end{aligned} \tag{4}$$

with $\mathbf{Y}(0) = \mathbf{X}(0)$. Given the fact that matrix $\mathbf{W}$ is doubly stochastic, we can see that $\frac{\mathbf{1}\mathbf{1}^\top}{n}\mathbf{X}(t) = \frac{\mathbf{1}\mathbf{1}^\top}{n}\mathbf{Y}(t) = \overline{\mathbf{X}}$, for all $t \geq 0$. In other words, Algorithm 1 maintains the mean of $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ constant.

$\diamond$ **Comparison:** Algorithm 1 implicitly yields the following three methods:

- **Exact Gossip (EG) [10]:** $\sigma = 0, \omega = 0$,
- **Compressed Gossip (CG) [21]:** $\sigma = 0, \omega \in [0, 1)$,
- **Scalable Exact Gossip (SEG) [12]:** $\sigma = \frac{5n - \sqrt{\gamma}}{5n + \sqrt{\gamma}}, \omega = 0$.

Note that *SEG* and *SCG* require the agents know the network size n or some $U = \mathcal{O}(n)$ to compute σ (see [12]).

Before stating the main results, let us compare our algorithm with prior works. Table 1 illustrates the linear convergence rates of the algorithms mentioned above along with a conservative bound for their feasible step-size γ and compression ratio ω. First, *EG* and *SEG* linear rates, which require exact communication, have a quadratic and linear dependence on n respectively. We will discuss in Section 3 how γ impacts the spectral gap of the mixing matrix in (4). Second, *CG* enjoys an arbitrary compression with a rate of $\mathcal{O}(n^4)$, but the choice of γ is limited to $\mathcal{O}(n^{-4})$. In this work, we use a different technique to analyze our algorithm *SCG*, where we restrict the choice of ω and let γ be arbitrary. As shown in Table 1, *CG* and *SCG* enjoy the same convergence rates as *EG* and *SEG*, with bounded γ. Given a reasonable bound for ω, our algorithm has a better dependence on n than *CG* given the same step-size γ. We conjecture that γ offers a trade-off between the convergence rate and the value of ω. In other words, with decreasing γ proportional to n^{-1}, the feasible set for ω expands proportional to n, which implies a worse convergence rate dependence on n. We will illustrate this trade-off in Fig. 2.

Table 1: Comparison of the worst case convergence rates for *EG*, *SEG*, *CG*, and *SCG*.

Algorithm	Linear Rate[a]	Stepsize (γ)	ω
EG [10]	$\mathcal{O}\big(1-\gamma n^{-2}\big)$	$(0, 1]$	0
SEG [12]	$\mathcal{O}\big(1-\gamma \frac{1}{2} n^{-1}\big)$	$\big(0, \frac{1}{2}\big]$	0
CG [21]	$\mathcal{O}\big(1-n^{-4}\big)$	$\mathcal{O}\big(n^{-4}\big)$	$[0, 1)$
CG[b]	$\mathcal{O}\big(1-\gamma n^{-2}\big)$	$(0, 1]$	$\Big[0, \Theta\Big(\frac{1}{(1+\gamma)n^2}\Big)\Big]$
SCG This Work	$\mathcal{O}\big(1-\gamma \frac{1}{2} n^{-1}\big)$	$\big(0, \frac{1}{2}\big]$	$\Big[0, \Theta\Big(\frac{1}{(1+\gamma)n^2}\Big)\Big]$[c]

[a] Convergence rates are linear, with different dependence on n and γ. Rates are presented for the worst case graphs where $\delta(\mathbf{W}) = \mathcal{O}(n^{-2})$.

[b] An alternative analysis for *CG* with bounded ω and flexible γ.

[c] Asymptotic bound for ω in Theorem 2.

3 Convergence Results

Here, we first propose the convergence guarantees for *SEG* and then *SCG*. As we mentioned earlier, under $\omega = 0$, (4) turns into the update rule for *SEG*:

$$\begin{aligned} \mathbf{Y}(t{+}1) &\coloneqq \mathbf{X}(t) + \gamma\big(\mathbf{W}{-}\mathbf{I}\big)\mathbf{X}(t), \\ \mathbf{X}(t{+}1) &\coloneqq (1{+}\sigma)\mathbf{Y}(t{+}1) - \sigma\mathbf{Y}(t), \end{aligned} \tag{5}$$

The following theorem states the convergence rate for (5).

Theorem 1 (An extension of Theorem 2.1 from [12])**.** *Let stepsize $\gamma \in \big(0, \frac{1}{2}\big]$, $\mathbf{Y}(0) = \mathbf{X}(0)$, and $\mathbf{W} = \mathcal{MH}(\mathcal{G})$. The following property holds for the update rule in (5):*

$$\Psi_x(t) \leq 2\lambda^t \Psi_x(0),$$

where $\Psi_x(t){=}\big\|\mathbf{X}(t){-}\overline{\mathbf{X}}\big\|_F$, $\lambda{=}1{-}\frac{\sqrt{\gamma}}{5n}$, when $\sigma{=}\frac{5n-\sqrt{\gamma}}{5n+\sqrt{\gamma}}$.

The result in Theorem 1 holds for an arbitrary stepsize $\gamma \in (0, 1/2]$, compared to [12] that holds for $\gamma = 1/2$ only. The auxiliary mixing matrix used by both *SEG* and *SCG* have the same dependence on γ, so the analysis for *SEG* helps to understand the analysis for *SCG* better.

Theorem 2 (SCG Convergence Analysis)**.** *Let compression operator Q satisfies (3), $\mathbf{Y}(0) = \mathbf{X}(0)$, $\hat{\mathbf{X}}(0) = \mathbf{0}$, and γ, σ, λ, and $\mathbf{W}$ be as Theorem 1. Then, the update rule in (4) satisfies the following: for $\omega \leq \big(2\big(\kappa_3 + \gamma\beta\kappa_2\big)\big(\lambda^{-\frac{1}{2}} + \gamma\beta\kappa_2 C\lambda^{-1}(1-\lambda^{\frac{1}{2}})^{-2}\big)\big)^{-1}$*

$$\mathbb{E}\Psi_x(t) \leq C_0\tilde{\lambda}^t\Psi_x(0),$$

where $\kappa_2{=}\sqrt{2\sigma^2{+}2\sigma{+}1}$, $\kappa_3 = \sqrt{2\sigma^2{+}2}$, $\beta = \|\mathbf{W}{-}\mathbf{I}\|$, $\tilde{\lambda}{=}1{-}\frac{\sqrt{\gamma}}{10n}$, $\Psi_x(t){=}\big\|\mathbf{X}(t){-}\overline{\mathbf{X}}\big\|_F$, and constants $C_0, C > 0$.

The above theorem implies a linear convergence for Algorithm 1 with rate $\tilde{\lambda}$ dependent on $\gamma^{-1/2}n$ under a bounded compression ratio ω, where the bound on ω can be written as $\Theta\big((1{+}\gamma)^{-1}n^{-2}\big)$.The above bound suggests that the consensus step-size γ imposes a trade-off between the convergence rate and the compression ratio ω.

Before stating the proofs, we propose a technical lemma that will help us prove Theorems 1 and 2.

Lemma 1. *Let matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric, doubly stochastic, and diagonally dominant with $\lambda_2(\mathbf{A}) \leq 1{-}\frac{1}{p^2}$, for some $p > 1$, and $\mathbf{B} \in \mathbb{R}^{2n\times 2n}$ be as follows:*

$$\mathbf{B} = \begin{bmatrix} (1{+}\sigma)\mathbf{A} & -\sigma\mathbf{A} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}.$$

Let $\lambda = 1{-}\frac{1}{p}$ and $\sigma = \frac{p-1}{p+1}$, then following statements hold:

(a) [12, Lemma 2.5] If $\mathbf{v} = [\mathbf{q}^\top, \mathbf{q}^\top]^\top$ and $\overline{\mathbf{v}} = [\overline{\mathbf{q}}^\top, \overline{\mathbf{q}}^\top]^\top$, for an arbitrary $\mathbf{q} \in \mathbb{R}^n$ with $\overline{\mathbf{q}} = \frac{\mathbf{1}\mathbf{1}^\top}{n}\mathbf{q}$, then $t \geq 0$,

$$\|\mathbf{B}^t\mathbf{v} - \overline{\mathbf{v}}\| \leq 2\lambda^t\|\mathbf{v} - \overline{\mathbf{v}}\|.$$

(b) If $\mathbf{v} = [\mathbf{q}^\top, \mathbf{0}^\top]^\top$, for $\mathbf{q} \in \mathbb{R}^n$ such that $\mathbf{1}^\top\mathbf{q} = 0$, then for all $t \geq 0$,

$$\|\mathbf{B}^t\mathbf{v}\| \leq Ct\lambda^t,$$

where $C > 0$ is some constant.

Proof sketch for Lemma 1. Similar to [12, Lemma 2.3], by considering the SVD-decomposition of $\mathbf{A}$, the problem reduces to show the convergence of $[\mathbf{B}(\lambda)]^t\mathbf{r}$, for $\mathbf{r} = [1, 1]^\top$, and $\mathbf{r} = [1, 0]^\top$, where $\mathbf{B}(\lambda_i) = \begin{bmatrix} (1{+}\sigma)\lambda_i & -\sigma\lambda_i \\ 1 & 0 \end{bmatrix}$ is a 2×2 matrix, for $i \in \{2, 3, \ldots, n\}$, and $1 - \lambda_i \geq p^{-2}$.

The analysis in [12] shows the convergence for $\mathbf{r} = [1, 1]^\top$, but their method is restricted to vectors $\mathbf{r}$ with the same two elements. However, this is not the case for Lemma 1(b), thus we consider an alternative technique. Note that $[\mathbf{B}(\lambda)]^t\mathbf{r}$ implies a recursive sequence with the following definition:

$$a(t) = (1+\sigma)\lambda\, a(t-1) - \sigma\lambda\, a(t-2), \quad \text{for all } t \geq 2, \tag{6}$$

with $a(1) = 1$, and either $a(0) = 0$ or 1. To find $a(t)$, we consider its corresponding generating function

$$G(x) = \frac{[a(1) - (1+\sigma)\lambda\, a(0)]\, x + a(0)}{\sigma\lambda\, x^2 - (1+\sigma)\lambda\, x + 1}, \tag{7}$$

where one can find the exact form of $a(t)$ given the choices for $a(1)$ and $a(0)$. The exact solution for $a(t)$ completes the proof for Lemma 1. □

We need Lemma 1 in the proof for both theorems, and Lemma 1(b) for Theorem 2. Next, we show the proof sketch for Theorem 1.

Proof sketch for Theorem 1. Let $\mathbf{M} = (1-\gamma)\mathbf{I} + \gamma\mathbf{W}$, be a lazy version of $\mathbf{W}$ defined in (2), thus $\mathbf{M}$ is also a doubly stochastic matrix with $\delta(\mathbf{M}) = \gamma\delta(\mathbf{W})$. We seek to derive a lower bound of $\mathcal{O}(\gamma/n^2)$ on the spectral gap of matrix $\mathbf{M}$. Our proof follows the structure of [12, Theorem 2.1], but we consider an arbitrary $\gamma \in (0, 1/2]$, which will also be used for Theorem 2. Note that a doubly stochastic matrix can be interpreted as a Markov chain's transition matrix. Now, assume that $\mathbf{M}$ is the transition matrix associated with a Markov chain. We know that $\mathbf{M}$ is a convex combination of $\mathbf{I}$ and $\mathbf{W}$, which implies with probability γ, the matrix $\mathbf{W}$ determines the transitions of the chain. Hence, using the result in [28], we can infer that the following property holds for the hitting time[1] of $\mathbf{M}$ [31]:

$$\max_{i,j\in[n]} \mathcal{H}_{\mathbf{M}}(i \to j) \leq \frac{6n^2}{\gamma}. \tag{8}$$

Moreover, by [31, Theorem 12.4 and Theorem 10.14],

$$\left(\frac{1}{\delta(\mathbf{M})} - 1\right) \ln 2 \leq 2 \max_{i,j\in[n]} \mathcal{H}_{\mathbf{M}}(i \to j) + 1, \tag{9}$$

so, due to (8) and (9), we have $\delta(\mathbf{M}) \geq \gamma/25n^2$. The rest of the proof is an immediate result of Lemma 1(a). □

We present the proof for Theorem 2 in Appendix A.

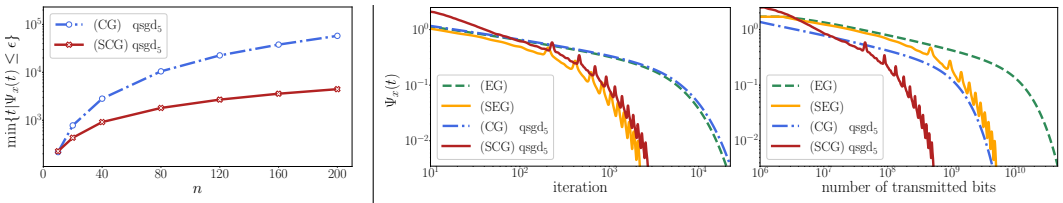
4 Numerical Experiments

Here, we present a set of numerical results to illustrate the communication advantages of our method. We consider the decentralized average consensus problem for a set of n agents with vectors of size $d = 150$. We consider two classes of networks with slow mixing times, path and ring, as well as operator qsgd_k for message compression.

Figure 1 presents two different experiments. First, we compare the performance of *CG* versus *SCG* given the same quantization operators, qsgd_5. We consider path graphs with size n varying from 10 to 200, and given a random set of initial parameters, consider the number of iterations t for each algorithm to reach an ϵ-consensus, i.e., $\Psi_x(t) \leq \epsilon$, for $\epsilon = 10^{-4}$. We run each algorithm 10 times and average the results. We apply a grid line search for the optimal γ in each case. As shown in Fig. 1a, our algorithm requires a fewer number of iterations to reach consensus compared to *CG*.

We furthermore provide a comparison between *EG*, *SEG*, *CG*, and *SCG* in Fig. 1b. We consider a ring graph with $n = 120$, and random parameters with dimension $d = 150$. We show the decay of $\Psi_x(t)$ based on the number of communications (left) as well as the number of transmitted bits (right). Figure 1b shows that *SCG* requires approximately the same number of communication rounds as *SEG*, with only 10% of bits transmitted to reach the same accuracy $\epsilon = 10^{-4}$.

[1] For a Markov Chain with transition matrix $\mathbf{W}$, hitting time $\mathcal{H}_{\mathbf{W}}(i \to j)$ indicates the expected number of steps for the chain to reach state j starting from state i.



(a) Path graphs with $n = 10$ to 200

(b) Ring network with size $n = 120$

Figure 1: **Scalability Numerical Analysis:** Each experiment is the average of 10 runs. **(a)** Comparison between the number of iteration required for *CG* and *SCG* to reach an ϵ-convergence on the average consensus problems with $d = 150$, for path networks with n ranging from 10 to 200, qsgd_5, and $\epsilon = 10^{-4}$. **(b)** Comparison of the ϵ-suboptimality for algorithms in Table 1, for an average consensus problem with $d = 150$, qsgd_5, $\epsilon = 10^{-4}$, over a ring graph with size $n = 120$ based on the number of iterations (left) and the number of transmitted bits (right).

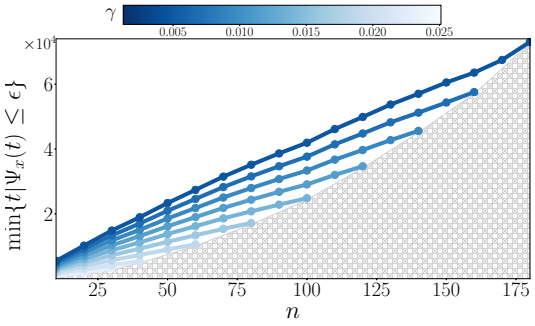


Figure 2: **Stepsize impact on convergence rate:** Each simulation is the average of 10 runs. Each point shows the number of rounds required for *SCG* to reach an ϵ-consensus ($\epsilon = 10^{-3}$) for an average consensus problem with $d = 100$ over a path network with n agents using the compression operator qsgd_3. Each line associates with a fixed stepsize γ, and the hatched area shows the set of n, which for *SCG* diverges given the corresponding γ.

We end this section with an example that explains the role of step-size γ in the trade-off between the convergence rate and compression feasibility. Similar to Fig. 1a, we consider the number of iterations for our algorithm to reach an ϵ-convergence for an average consensus problem with $d = 100$, over path networks with varying size n with quantizer qsgd_3. We consider a range of step-sizes $\gamma \in [0.001, 0.025]$, and for each one, we run our algorithm for different choices of n. As shown in Figure 2, given a fixed quantization ratio, γ imposes a trade-off between the convergence rate versus the feasibility of the consensus for ω. Hence, a better rate requires a larger γ, which requires a smaller compression ratio ω, while for a larger ω, we need to decrease γ, which slows down the convergence.

5 Conclusions

In this work, we proposed a scalable communication-efficient algorithm for the problem of decentralized average consensus. Given a large enough compression ratio, we showed that agents can communicate compressed messages yet reach consensus with a linear rate that depends linearly on the number of agents in the network. We further presented numerical results to illustrate our theoretical studies. Future work should investigate the combined effect of communication efficiency and scalability in decentralized problems like optimization and inference using the proposed consensus technique. The impact of byzantine agents and other variations of the consensus problem remain as future work.

References

[1] A. Nedić, A. Olshevsky, and C. Uribe, "Fast Convergence Rates for Distributed Non-Bayesian Learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

[2] A. Nedić and A. Ozdaglar, "Distributed Subgradient Methods for Multi-Agent Optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48–61, 2009.

[3] A. Nedić, A. Ozdaglar, and P. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[4] A. Lalitha, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

[5] E. Manley, H. Al Nahas, and J. Deogun, "Localization and tracking in sensor systems," in *IEEE International conference on sensor networks, ubiquitous, and trustworthy computing (SUTC'06)*. IEEE, 2006, vol. 2, pp. 237–242.

[6] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.* IEEE, 2005, pp. 63–70.

[7] A. Syed, J. Heidemann, et al., "Time Synchronization for High Latency Acoustic Networks.," in *Infocom*, 2006, vol. 6, pp. 1–12.

[8] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[9] J. Wang, Z. Charles, Z. Xu, G. Joshi, H.B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al., "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[10] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[11] K. Cai and H. Ishii, "Average consensus on arbitrary strongly connected digraphs with time-varying topologies," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 1066–1071, 2014.

[12] A. Olshevsky, "Linear Time Average Consensus and Distributed Optimization on Fixed Graphs," *SIAM J. Control. Optim.*, vol. 55, pp. 3990–4014, 2017.

[13] A. Kashyap, T. Başar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.

[14] A. Nedić, A. Olshevsky, A. Ozdaglar, and J.N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on automatic control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[15] P. Frasca, R. Carli, F. Fagnani, and S. Zampieri, "Average consensus on networks with quantized communication," *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, vol. 19, no. 16, pp. 1787–1816, 2009.

[16] G. Baldan and S. Zampieri, "An efficient quantization algorithm for solving average-consensus problems," in *2009 European Control Conference (ECC)*. IEEE, 2009, pp. 761–766.

[17] D. Thanou, E. Kokiopoulou, Y. Pu, and P. Frossard, "Distributed average consensus with quantization refinement," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 194–205, 2012.

[18] T. Can Aysal, M. Coates, and M. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4905–4918, 2008.

[19] R. Carli, F. Bullo, and S. Zampieri, "Quantized average consensus via dynamic coding/decoding schemes," *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, vol. 20, no. 2, pp. 156–175, 2010.

[20] K. Cai and H. Ishii, "Quantized consensus and averaging on gossip digraphs," *IEEE Transactions on Automatic Control*, vol. 56, no. 9, pp. 2087–2100, 2011.

[21] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication," in *International Conference on Machine Learning*, 2019, pp. 3478–3487.

[22] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtarik, and S.U. Stich, "A linearly convergent algorithm for decentralized optimization: Sending less bits for free!," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 4087–4095.

[23] Hossein Taheri, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9324–9333.

[24] M.T. Toghani and C. Uribe, "Communication-efficient distributed cooperative learning with compressed beliefs," *arXiv preprint arXiv:2102.07767*, 2021.

[25] Z. Song, L. Shi, S. Pu, and M. Yan, "Compressed gradient tracking for decentralized optimization over general directed networks," *arXiv preprint arXiv:2106.07243*, 2021.

[26] A. Nedić, A. Olshevsky, and C. Uribe, "Graph-theoretic analysis of belief system dynamics under logic constraints," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.

[27] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[28] Y. Nonaka, H. Ono, K. Sadakane, and M. Yamashita, "The hitting and cover times of metropolis walks," *Theoretical Computer Science*, vol. 411, no. 16-18, pp. 1889–1894, 2010.

[29] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[30] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, "On biased compression for distributed learning," *arXiv preprint arXiv:2002.12410*, 2020.

[31] D. Levin and Y. Peres, *Markov chains and mixing times*, vol. 107, American Mathematical Soc., 2017.

Appendix

A Proof of Theorem 2

Proof for Theorem 2. Let us define 2×2 matrices $\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{T}_3$ as follows:

$$\mathbf{T}_1 = \begin{bmatrix} 1+\sigma & -\sigma \\ 1 & 0 \end{bmatrix}, \tag{10}$$

$$\mathbf{T}_2 = \begin{bmatrix} 1+\sigma & -\sigma \\ 0 & 0 \end{bmatrix}, \qquad \kappa_2 = \|\mathbf{T}_2\| = \sqrt{2\sigma^2+2\sigma+1},$$

$$\mathbf{T}_3 = \begin{bmatrix} \sigma & -\sigma \\ 1 & -1 \end{bmatrix}, \qquad \kappa_3 = \|\mathbf{T}_3\| = \sqrt{2\sigma^2+2}.$$

We furthermore define $\mathbf{Z}(t)$, $\hat{\mathbf{Z}}(t)$, and $\overline{\mathbf{Z}}$, as

$$\mathbf{Z}(t) = \begin{bmatrix} \mathbf{X}(t) \\ \mathbf{X}(t-1) \end{bmatrix}, \ \hat{\mathbf{Z}}(t) = \begin{bmatrix} \hat{\mathbf{X}}(t) \\ \hat{\mathbf{X}}(t-1) \end{bmatrix}, \ \overline{\mathbf{Z}} = \begin{bmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{X}} \end{bmatrix}, \tag{11}$$

with initialization $\hat{\mathbf{X}}(0) = \mathbf{0}$ and $\mathbf{X}(-1) = \mathbf{X}(0)$. Therefore, the update rule in (4) can be rewritten as follows:

$$\begin{aligned} \mathbf{Z}(t+1) &= [\mathbf{T}_1 \otimes \mathbf{I}]\mathbf{Z}(t) + \gamma[\mathbf{T}_2 \otimes (\mathbf{W}-\mathbf{I})]\hat{\mathbf{Z}}(t+1) \\ &= \mathbf{B}\mathbf{Z}(t) + \gamma[\mathbf{T}_2 \otimes (\mathbf{W}-\mathbf{I})]\big(\hat{\mathbf{Z}}(t+1)-\mathbf{Z}(t)\big) \\ &= \mathbf{B}^{t+1}\mathbf{Z}(0)+\gamma\sum_{s=0}^{t}\mathbf{B}^s[\mathbf{T}_2\otimes(\mathbf{W}-\mathbf{I})](\hat{\mathbf{Z}}(t-s+1)-\mathbf{Z}(t-s)). \end{aligned} \tag{12}$$

We now define the Lyapunov functions $\mathcal{R}_z(t)$ and $\mathcal{U}_z(t)$ as

$$\mathcal{R}_z(t) \triangleq \big\|\mathbf{Z}(t) - \overline{\mathbf{Z}}\big\|_F, \quad \mathcal{U}_z(t) \triangleq \big\|\hat{\mathbf{Z}}(t+1) - \mathbf{Z}(t)\big\|_F,$$

and bound them using Lemma 1. First, we have

$$\begin{aligned} \big\|\mathbf{Z}(t+1) - \overline{\mathbf{Z}}\big\|_F &\overset{\text{tri. ineq.}}{\leq} \big\|\mathbf{B}^{t+1}\mathbf{Z}(0)-\overline{\mathbf{Z}}\big\| + \gamma\sum_{s=0}^{t}\big\|\mathbf{B}^s[\mathbf{T}_2 \otimes (\mathbf{W}-\mathbf{I})]\big(\hat{\mathbf{Z}}(t-s+1) - \mathbf{Z}(t-s)\big)\big\|_F \\ &\overset{\text{Lemma 1(a), (b)}}{\leq} 2\lambda^{t+1}\big\|\mathbf{Z}(0)\big\|_F + \gamma\beta\kappa_2 C\sum_{s=0}^{t} s\lambda^s\big\|\hat{\mathbf{Z}}(t-s+1) - \mathbf{Z}(t-s)\big\|_F. \end{aligned} \tag{13}$$

Using the definition of $\hat{\mathbf{X}}(t)$ in (4), we also have

$$\begin{aligned} \mathbb{E}\big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+2)\big\|_F^2 &\overset{(11)}{=} \mathbb{E}\big\|\mathbf{X}(t+1) - \hat{\mathbf{X}}(t+2)\big\|_F^2 + \mathbb{E}\big\|\mathbf{X}(t) - \hat{\mathbf{X}}(t+1)\big\|_F^2 \\ &\overset{(3)}{\leq} \omega^2\left[\big\|\mathbf{X}(t+1) - \hat{\mathbf{X}}(t+1)\big\|_F^2 + \big\|\mathbf{X}(t) - \hat{\mathbf{X}}(t)\big\|_F^2\right] \\ &= \omega^2\big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+1)\big\|_F^2, \end{aligned} \tag{14}$$

where according to Jensen's inequality, and (14) we have

$$\mathbb{E}\big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+2)\big\|_F \leq \omega\big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+1)\big\|_F. \tag{15}$$

Hence, we need to bound $\big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+1)\big\|_F$, as follows:

$$\begin{aligned} \big\|\mathbf{Z}(t+1) - \hat{\mathbf{Z}}(t+1)\big\|_F &\overset{(12)}{=} \big\|[\mathbf{T}_1 \otimes \mathbf{I}]\mathbf{Z}(t) + \gamma[\mathbf{T}_2 \otimes (\mathbf{W}-\mathbf{I})]\hat{\mathbf{Z}}(t+1) - \hat{\mathbf{Z}}(t+1)\big\|_F \\ &= \big\|[\mathbf{I}_{2n} + \gamma\mathbf{T}_2 \otimes (\mathbf{I} - \mathbf{W})]\left(\mathbf{Z}(t) - \hat{\mathbf{Z}}(t+1)\right) \\ &\quad + [\mathbf{T}_3 \otimes \mathbf{I}_n + \gamma T_2 \otimes (\mathbf{W}-\mathbf{I}_n)]\left(\mathbf{Z}(t) - \overline{\mathbf{Z}}\right)\big\|_F \\ &\overset{(10)}{\leq} (1 + \gamma\beta\kappa_2)\big\|\mathbf{Z}(t) - \hat{\mathbf{Z}}(t+1)\big\|_F \\ &\quad + (\kappa_3 + \gamma\beta\kappa_2)\big\|\mathbf{Z}(t) - \overline{\mathbf{Z}}\big\|_F. \end{aligned} \tag{16}$$

Based on (13), (14), and (15), we have:

$$\mathbb{E}\mathcal{R}_z(t{+}1) \leq 2\lambda^{t+1}\big\|\mathbf{Z}(0)\big\|_F + \gamma\beta\kappa_2 C\sum_{s=0}^{t} s\lambda^s\,\mathcal{U}_z(t{-}s),$$
$$\mathbb{E}\,\mathcal{U}_z(t{+}1) \leq \omega(1{+}\gamma\beta\kappa_2)\,\mathcal{U}_z(t) + \omega(\kappa_3{+}\gamma\beta\kappa_2)\,\mathcal{R}_z(t).$$

Let $\nu = \omega(\kappa_3 + \gamma\beta\kappa_2)$, $\nu' = \omega(1 + \gamma\beta\kappa_2)$, where $\nu \geq \nu'$. We now by induction show that for

$$\omega \leq \frac{1}{2(\kappa_3 + \gamma\beta\kappa_2)(\lambda^{-\frac{1}{2}} + \gamma\beta\kappa_2 C\lambda^{-1}(1-\lambda^{\frac{1}{2}})^{-2})}, \tag{17}$$

$\mathcal{U}_z(t)$ satisfies the following inequality:

$$\mathcal{U}_z(t) \leq \xi_0\lambda^{t/2}, \tag{18}$$

where $\xi_0 = 4\lambda^{-\frac{1}{2}}\nu\big\|\mathbf{Z}(0)\big\|_F$. First, one can check (18) holds for $t = 0$. Furthermore,

$$\begin{aligned}
&\mathbb{E}\,\mathcal{U}_z(t{+}1)\\
&\leq \nu\,\mathcal{U}_z(t) + 2\nu\lambda^t\big\|\mathbf{Z}(0)\big\|_F + \gamma\nu\beta\kappa_2 C\sum_{s=0}^{t-1} s\lambda^s\mathcal{U}_z(t{-}s{-}1)\\
&\leq \nu\,\xi_0\lambda^{\frac{t}{2}} + 2\nu\lambda^t\big\|\mathbf{Z}(0)\big\|_F + \gamma\nu\beta\kappa_2 C\xi_0\sum_{s=0}^{t-1} s\lambda^s\lambda^{\frac{t-s-1}{2}}\\
&\leq \left(\frac{2\nu}{\lambda^{\frac{1}{2}}}\big\|\mathbf{Z}(0)\big\|_F + \frac{\nu\xi_0}{\lambda^{\frac{1}{2}}} + \frac{\gamma\nu\beta\kappa_2 C\xi_0}{\lambda(1-\lambda^{\frac{1}{2}})^2}\right)\lambda^{\frac{t+1}{2}}\\
&\leq \left(\frac{\xi_0}{2} + \frac{\xi_0}{2}\right)\lambda^{\frac{t+1}{2}} \leq \xi_0\lambda^{\frac{t+1}{2}},
\end{aligned} \tag{19}$$

where we used $\sum_{s=0}^{\infty} s\lambda^{\frac{s}{2}} \leq \left(1-\lambda^{\frac{1}{2}}\right)^{-2}$, using its corresponding generating function. We then bound $\mathcal{R}_z(t)$:

$$\begin{aligned}
\mathbb{E}\mathcal{R}_z(t) &\leq 2\lambda^t\big\|\mathbf{Z}(0)\big\|_F + \gamma\beta\kappa_2 C\sum_{s=0}^{t-1} s\lambda^s\,\mathcal{U}_z(t{-}s{-}1)\\
&\leq \frac{\xi_0}{2\nu}\lambda^{t-1} + \frac{\gamma\beta\kappa_2 C}{\lambda^{\frac{1}{2}}}\sum_{s=0}^{t-1} s\lambda^s\,\mathcal{U}_z(t{-}s{-}1)\\
&\leq \underbrace{\xi_0\left(\frac{\lambda^{\frac{t}{2}}}{2\nu\lambda} + \frac{\gamma\beta\kappa_2 C}{\lambda\left(1-\lambda^{\frac{1}{2}}\right)^2}\right)}_{C_0\text{: constant}}\lambda^{\frac{t}{2}} = C_0\,\lambda^{\frac{t}{2}}.
\end{aligned} \tag{20}$$

We moreover know that

$$\sqrt{\lambda} \leq \sqrt{1 - \frac{\sqrt{\gamma}}{5n} + \frac{\gamma}{100n^2}} = 1 - \frac{\sqrt{\gamma}}{10n} = \tilde{\lambda}, \tag{21}$$

then, $\mathcal{R}_z(t) \leq C_0\tilde{\lambda}^t$, which concludes the proof. □